uc3m | Universidad Carlos III de Madrid

Grado Universitario de Ingeniería Informática

2018-2019

Trabajo Fin de Grado

"Minería de texto en redes sociales"

Borja Varas Chavero

Tutor/es

José Antonio Iglesias Martínez

Leganés, Junio de 2019

Esta obra se encuentra sujeta a la licencia Creative Commons **Reconocimiento – No Comercial – Sin Obra Derivada**

Agradecimientos

El presente trabajo de fin de carrera es el final de un largo y bonito camino que ha durado cinco años, en los cuales esta universidad me ha visto crecer, madurar y aprender, tanto personalmente como académicamente.

Por eso, me gustaría agradecerle a la universidad el hecho de formarme en todos los ámbitos de la forma que lo ha hecho.

Gracias a todos los profesores que se han cruzado en esta carrera de fondo por ayudarme y enseñarme este campo, como es la informática, desde otras perspectivas que para mí eran desconocidas.

Gracias a mi tutor, José Antonio, por ayudarme con tantas ganas y atención como ha hecho desde el primer momento en el que comenzamos a trabajar en este trabajo.

Gracias a los compañeros con los que he trabajado en estos años, con los que he aprendido a trabajar en grupo.

Agradecerles a mis amigos y mi familia el apoyo en cada momento que me lo han brindado.

Para finalizar, agradecerles a mis padres, haber hecho posible que yo haya llegado hasta aquí, por darme la confianza y la fuerza, cuando estas estaban ausentes en mí para continuar esta bonita y dura etapa de mi vida.

Abstract

INTRODUCTION AND MOTIVATION

Nowadays, we have access to a large amount of data. This is now possible thanks to the technology and the tools that have been develop, which offer new opportunities to society, both personally and at the company level.

Our behaviors in social networks, such as: hashtags that we use, searches we make, likes we give to publications, or retweets; create a digital mark or trail that facilitate how the user is, their concerns, interests, hobbies or desires. The current way of analyzing all those data that users contribute without realizing is the sentiment analysis. Through this tool you can analyze the opinions or impressions of a user on a specific topic, a service, a product or your political preference, as will be seen later on and that will be analyzed in this work.

On the other hand, so much data and information must be processed, grouped and classified. Here the automatic classification and grouping processes come into play.

The concept of Big Data, and this massive generation of data that exists in a global way, develops new methods and techniques to perform a correct analysis, management and storage of data. Increasing interest is shown by the data and its analysis for extraction of useful information, given that this analysis and treatment is fundamental to make predictions, detect patterns or help companies make decisions.

These techniques have taken on greater importance in society at all levels in past years.

The text is one of the most used ways to formalize the data. Added to this is the fact that the largest sources of data today are large companies and social networks, since they are platforms that operate globally and internationally.

Text mining is one of the branches of computational linguistics that seeks to obtain information and knowledge from data sets that, in principle, do not have an order or are not ready at origin to transmit that information.

In addition, it is a key technique in a world like the one in which data are continuously collected from different perspectives and from many different aspects of all the activities of human beings.

Text mining has three fundamental activities such as:

- Recovery of information, that is, select the relevant texts.
- Extraction of the information included in these texts: facts, events, key data, relationships between them.

- Finally, we would do what we defined before as data mining to find associations between those key data previously extracted from the texts.

Many of the data generated by these entities are not analyzed and therefore lose a lot of useful information that would allow them to evolve and improve.

OBJECTVES

The principal objectives of this work are the following:

- Collection of tweets for the subsequent implementation of a confidence classification model. With a trust model, it can be grouped according to the theme of the text to get better consultations.
- Extraction of useful information and establishment of relationships between classification and grouping data.
- Analysis of the texts of the users of Twitter before political actions in order to extract the opinion about the political parties.
- Use sentiment analysis to forecast voting results such as polls or surveys, that is, show that this type of analysis can be valid tool like the ones mentioned above.
- Relate the political events that have taken place in a specific period of time with the analysis made on the polarity of the tweets taken from Twitter.

STATE OF ART

OBTAINING DATA AND GENERATION OF FILES

To select the data sets that were part of this experimentation, several datasets were taken from the network. As the purpose of the data collection is to generate sets of reliable and real instances, and these showed errors, noise and completeness problems in their instances, it was opted for the own collection of a new dataset.

In addition to the problems presented by these datasets at the data level, they were not adapted to the needs of experimentation. The formula by which we opted to generate the new data set is the obtaining of the data through the Twitter API. Through the API it was possible to recover tweets of the users filtering through different fields, which, after all, is what was sought. The tool used to collect tweets was Twitter Archiver.

Twitter Archiver is an extension of Google that allows you to save queries made on Twitter directly in a Google Spreadsheet. This tool allows you to perform different filters in the search of texts, such as searching for who has written it, whether the text is retweeted or not, how many retweets it has, by location or by hashtags.

The search for tweets was done using the hashtags “#Policy”, “#Economy”, “#Technology” and “#Sports”, given that they are the topics on which we want to obtain, classify and group the tweets.

The first problem that arises is that many tweets are needed to carry out the study and with the free version of Twitter Archiver you can collect 100 per hour at the most. To this is added that, when made through a Google account, only one query is allowed per account simultaneously, that is, you can only collect tweets of a topic.

In order for the collection task to take no longer than desired, four Google accounts were created in which the Twitter Archiver extension was installed. In each of the accounts, the search for a topic was implemented, in this way, tweets of all the themes were collected simultaneously, shortening the time cost of this task.

Once enough tweets were obtained for our proposal (we need to consider that the tweets are not information, they are text data) a process was carried out in which those fields on the tweet that were not interesting for the study were eliminated. It was estimated that the only important and relevant data was the tweet itself, the text, so, all other fields were eliminated.

In addition, another field was added to the tweet field with the class or theme to which it belonged. In this way, there were four master spreadsheets with the tweets compiled by themes. This is important for the formation of the files with which it will be experimented later, since it facilitates a balanced composition of the classes, the randomization and manipulation of the tweets.

In relation to the data files collected for clustering tasks, they are the same as the files for classification, with the only difference that the class of the a priori text is eliminated. This elimination is due to the type of learning to which this task belongs.

In the case of sentiment analysis, the tweets are downloaded directly from RapidMiner, with the Search Twitter operator, making a query with the filters that the user determines.

ANALYSIS AND SYSTEM DESIGN

In this section, the functionalities of the system are described. This must meet a series of requirements, both functional and non-functional. In addition, some cases of use and the architecture of the system have been developed.

Obtaining requirements is an essential task to be able to define exactly the capabilities of the system. In addition, the requirements are a reference to verify the design. The requirements are classified in two categories:

- Functional: they are the ones that establish the technical part of the system, that is, they describe the functionality of the system.

- Non-functional: they describe the characteristics and limit how the software development should be carried out, that is, how the functional requirements are carried out.

Among the functional requirements can be highlighted some as those referred to: the obtaining of tweets from the social network, selection of the type of task to be performed, choice of task within the type of learning, choice of the type of classifier, selection of the data set to use, establishment of keys to perform a query to Twitter, selection of the attribute to be predicted or number of groups that are desired.

Regarding the non-functional requirements, some stand out, such as the status of the process that is being executed, the format of the data files, the structure of the classification and grouping files, connection with external tools or permissions between tools.

A use case consists of a description of the steps followed by an actor, that is, a user of the system, to carry out a task in the system.

Normally in the cases of use, the types of actors / users that perform the action are usually differentiated, but in this case, there would only be one possible actor, the user. For this reason, in these tables where the use cases are described, the type of user is not described.

The detected cases of use are those corresponding to making a query to Twitter, the generation of classification models from a classifier, the generation of a grouping model and the generation of sentiment analysis.

The architecture of the system shows how the different components that make up the system are related. The system consists of three main subsystems, which are described below.

- Subsystem for obtaining tweets. This subsystem is responsible for the task of compiling tweets from the user's Twitter account for later use in the different available actions. Different components can be distinguished depending on the destination of the tweets.

- File manager subsystem. This subsystem is responsible for the operations related to the data files and the preprocessing of them, more specifically, with the elimination of the hashtags of the texts. Different components are differentiated.

- Model generator subsystem. This subsystem is responsible for generating the models corresponding to the tasks available. It has a component for each of the tasks.

RESULTS

CLASSIFICATION

Supervised learning is a type of learning that is based on discovering the relationship between some input variables and some output variables. Learning arises by showing this type of algorithm what is the result you want to obtain for a certain value.

- **Classification cross validation**

  Cross-validation is a method or technique for evaluating an analytical process that aims to ensure that the results are independent for the distribution of test and training data. It consists in obtaining two separate sets staring from an original one, the first one is used for the training process and the second one for the validation process.

  Experiments will be carried out with different types of classification algorithms.

  - Naïve Bayes: it is one of the most used classifiers for its simplicity and speed. Its defined as a supervised classification and prediction technique that constructs models that define the probability of possible outcomes. It is based on the Bayes Theorem.

  - KNN is a lazy algorithm, which means that, during training, it only saves instances, does not build any model, not like decision trees or other algorithms, only performs the classification when it arrives at the test instances.

  - Deep Learning is based on an artificial neural network of advanced feeding that contains multiple layers and whose training is carried out by means of the descent of stochastic weight gradient using backward propagation. The network can be composed of numerous hidden layers formed by neurons with different activation functions.

  - A decision tree is a prediction model or a technique that allows analyzing sequential decisions based on the use of associated results and probabilities.

  The results obtained are the following:

  - Naive Bayes

    - In the first set, a 72.25% success is obtained, which is the worst result of this classification algorithm.

    - With the second set you get 80.74% success, which translates as the best result of the algorithm.

- In the third set you get a 76.22% success and in the fourth set a 78.33%.

- KNN

  With this algorithm, it was only possible to develop the process corresponding to the combination of this algorithm with the first set of instances. This process showed a success rate of 72.10%.

- Deep Learning

  With this algorithm, only the process corresponding to the first data set can be developed. obtaining a 77.90% success.

- Decision Tree

  With this algorithm it was possible to carry out three of the four experiments or processes that were planned. These correspond to the first three sets of instances and the results are the following:

  - With the first set a 36.88% accuracy was obtained.
  - For the second set of obtained a 33.75% success.
  - And for the third of the sets a 32.98% accuracy.

The classifier that has presented the best results is Naïve Bayes, reaching an average maximum value of 80.74% in the second set of his experiments.

The worst classifier has been the decision tree, presenting success values comparable to those presented by random classifiers. We can highlight the predictions of several classes with 0% accuracy.

By increasing the number of instances per set, the increase in the accuracy of the classifiers was sought, but the result obtained did not match expectations. In fact, in some cases the opposite effect is obtained.

- **Classification**

For the classification, the best models of the experiment carried out with the cross-validation, Naïve Bayes and Deep Learning have been chosen.

o Naïve Bayes

With this algorithm, all the planned experiments could be carried out satisfactorily.

- The experimentation with the first data set gives a 98.42% accuracy, the best result of all obtained.

- With the second set you get 97.04% success.

- In the third of the sets you get 95.45%.

- Finally, with the fourth set you get 93.51%.

As can be seen in the results obtained, by increasing the number of instances of the data sets, the results worsen by a small percentage. This fact, in this particular case, does not affect to appreciate that the results are very positive, but as it has been commented previously, the opposite effect has been obtained that was sought with that increase of the volume of input data, since the objective of this action was to improve the percentage of success of the experiments.

o Deep Learning

With Deep Learning, only three of the four experiments have been carried out. The processes performed correspond to the first three data sets.

- With the first set you get a 98.20% accuracy.

- With the second of the sets, 87.02% accuracy was obtained. This experiment presents a great decrease in the success with respect to the previous one.

- The third set has 82.79% accuracy.

In the particular case of this algorithm, the fact of increasing the volume of instances per set has had a very visible and negative effect. Likewise, in Naive Bayes, the worsening was not significant and from one set to another there was a 2% drop, in Deep Learning, from the first to the second there is a decrease of 11%. This fact will also be discussed below.

- In the case of Naïve Bayes, studying the particular rates of each class it is observed that, the decrease of the general rate is almost exclusively due to the decrease of the Economy rate. This conclusion is reached because the rate of the other classes, in spite of increasing the examples, almost do not change, only the Economy one varies more notably. It must be said that, the decline in this classifier is not very important, it still present very positive results.

- In the case of Deep Learning, the decrease in the general rate evolves differently than in Naïve Bayes. In this case the general decline if it is due to a worsening of almost all the particular rates. Except for sports, the other classes get worse from one study to another. This would explain why the

decrease in the general rate is more important since from 97.9% it goes to 82.79%. 15% less, when in Bayes it was reduced to penalties by 5%.

Both classifiers would be suitable for an adequate classification, but due to the evolution of the error and the results obtained, the best model for the prediction system is Naïve Bayes.

CLUSTER

This type of learning is the one that manages to produce knowledge only from the data that is taken as input, without explaining to the system what result one wants to obtain.

This learning looks for patterns of similarity between the input data, that is, it one thing is similar to another.

Grouping is one of the most important and used tasks of unsupervised learning.

This task is done through the K-means method. Said method starts assuming that the number of groups or clusters in which the instances have to be grouped is known. Its objective is to find the “best” allocation of points for different groups. The term “best” refers to maximizing inter-cluster distances and minimizing intra-cluster distances.

These results make it clear that regardless of the input file that has as parameter, the resulting grouping is practically the same. It was expected that, by increasing the number of instances from one set to another, the behavior of the grouping would improve from one experiment to another. This fact does not end surprising, given that, in the task of classification also expected and improvement between sets, but the opposite effect occurred.

With the above, it can be concluded that the grouping behavior, in this case, is independent of the input files.

In addition, knowing in advance that two classes present a relationship that confuses the system, one could conclude that the fact that there is a cluster that hosts almost all instances of the set would be related to this fact.

SENTIMENT ANALYSIS

To analyze the sentiment of Twitter users, the Rapid Miner tool has been used, complemented by the AYLIEN extension. Rapid Miner is a tool designed for data mining and analysis. Through its graphic environment allows to design or develop data analysis processes. It is developed in Java language and is multiplatform.

On the one hand, you have data collection through access to external data from RapidMiner, in this case when it comes to Twitter, and so that you can download

the desired tweets you have to provide an account of that social network and grant you the appropriate permissions Once this is done, the connection with the social network is configured, which facilitates the data of the connection.

On the other hand, and in order to obtain the feeling of a certain tweet, the tweets downloaded by RapidMiner are passed as input to the AYLIEN plugin. This complement allows obtaining an output that generates the "type of feeling" that the user reflects with the written tweet. This type can have three possible values: positive, negative or neutral.

The objective of the following experiment will be to obtain information about the opinion of users of the social network Twitter about the political parties of our country. This information obtained will be analyzed in detail. This experiment has been carried out in a specific period f time, but what is really wanted is that it can be applied at any other time.

The data has been obtained during the months of December and January 2018, randomly and as allowed by the tool itself. It should also be noted that the tool has presented several problems, such as that, from a number of executions, was not able to get more tweets or even not specify if they were positive, negative or neutral. However, these types of problems have been solved by obtaining the desired results.

The party that has received more positive impressions has been the Popular Party (Partido Popular/PP), with 69.11%, besides being the one with the lowest percentage of negative tweets, 27.39%. This yields two important data, the first that the results presented by the data are coherent and the other is the low relevance or importance of the opinions considered as neutral, since they are very few. This data or phenomenon will be discussed again later.

On the other hand, the party with the highest negative percentage is the Spanish Socialist Workers Party (Partido Socialista Obrero Español/PSOE)), 31.40%, which describes an opposite evolution to the PP, since it is also the least positive, with 63.28%.

In addition, the party with greater indecision or, what is the same, with a greater number of tweets classifies as neutral is also PSOE, with 5.3%.

As a conclusion, it could be obtained that, taking into account the analysis of the graphs, the right-wing parties have obtained more positive results than the left-wing parties in the collection period. This fact could be explained by the change of government in Andalucia, where PP, Ciudadanos and VOX started to govern, which is a historical fact, since, during the last 36 years PSOE and the left-wing parties governed. Also, we need to take into account the *bad general moment* in

which they are currently going through, or in recent months, both Unidas Podemos and PSOE.

Several news items have been linked to the results obtained, such as the exit of Íñigo Errejón from Unidas Podemos, the recognition of Guaido as the president of Venezuela or the proposals of Santiago Abascal, among others.

By analyzing the sentiment of the users of a social network, other studies could be carried out to understand the behavior of society. A part of the political field can be taken to other areas, where the opinion of the people who from society is key, as for example, see what to spend public money, what are the priorities of people according to their age and sex or see how global actions affect society itself.

FUTURE WORKS

With everything developed in this project and previously studied to carry out this work have been raised different studies and work that could be done as a continuation of this.

The main differences in the face of future work would be:

- The number of instances that form each file, adding more instances to them.
- The teams with which to carry out the studies.

Some examples of future work are the following:

- Carry out the cross validation with the same classifiers that have been used in the present work with other equipment and other sets of instances. By using other equipment with greater power, and files with more instances, it would pretend to obtained better results with classifiers.
- Repeat the grouping experiment with other data sets and other equipment in order to achieve more conclusive results than those obtained in this work.
- Conduct s study similar to that carried out in a period of general pre-elections, in order to analyze the opinion of society in a political event of great importance.
- Study the age ranges and geographical situations of the users based on their opinions on a particular political party or character.

CONCLUSIONS

In the study that has been carried out with cross-validation, four different classifiers (Naïve Bayes, Deep Learning, KNN and Decision Tree), were analyzed by cross-validation, where significant results were obtained. KNN and Decision Tree were left behind, given the results they showed were not positive enough to move on to the next

evaluation. Naïve Bayes and Deep Learning, on the other hand, went on to a second study where the models were applied to see their evolution in a direct classification. Both obtained very consistent and positive results, but the best classifier was Naïve Bayes.

It was expected that increasing the number of instances from one data set to another would have a positive effect on the learning process of the classifiers, but the opposite effect was obtained in most cases.

The second of the studies consisted of performing a grouping by means of unsupervised learning with the K-means algorithm with the same data sets with which the classification was made.

The results were not as good as expected, since a grouping was expected with four relatively homogeneous clusters in terms of the number of instances involved. Instead, a cluster was obtained that covered the vast majority of tweets of the sets and three groups almost empty, with hardly any instances.

The third and last study is related to the analysis of sentiment of users´ tweets regarding the political situation in the country.

The polarity of the texts was studied, classifying them as positive, negative and neutral. The results obtained were studied in a global way, seeing the evolution of the opinions of the users over time, collecting behavioral changes or striking data, as well as with each game.

To conclude if the proposed analysis could provide the information presented, those outstanding data and changes of opinion among the users were attempted to link to political events that occurred on the same day of the data collection or following days thereafter. When carrying out this linking task, it is concluded that, in most cases, those changes of opinion among the users or highlighted data had the backing of an event or important political news.

Resumen

En este trabajo de fin de grado se han realizado varios estudios y experimentos relacionados con la minería de dados en las redes sociales, más concretamente sobre Twitter. La minería de texto, a día de hoy, es una de las herramientas más importantes y que más está creciendo en el ámbito de extracción y análisis de información.

La extracción de información útil de los datos es clave en muchos aspectos, pero los más destacables y por los que está sufriendo una gran evolución es por su aportación al sector empresarial. El análisis de los datos generados por las empresas, tanto en su producción como en el funcionamiento interno, ayudan a estas a tomar decisiones sobre la mejora interna y, su fijación y consecución de objetivos.

Respecto a la minería de texto en las redes sociales, mucha de la información que circula en el entorno actual de la sociedad proviene de las redes sociales, son una fuente incombustible generando información, más concretamente textos. Estos textos generados por los usuarios de las redes sociales muestran de forma subjetiva los temas de actualidad o de mayor relevancia del momento. Por esto mismo, analizar dichas opiniones puede dar información importante sobre cómo es la sociedad.

Con el fin de estudiar la minería de datos en diferentes ámbitos, este trabajo se ha dividido en diferentes tareas. Primeramente, se implementó un clasificador de textos de confianza, realizando experimentos con diferentes algoritmos clasificadores y conjuntos de instancias. Los conjuntos de datos están formados por instancias de cuatro temas generales como son: la política, el deporte, la tecnología y la economía. Además, con estos mismos conjuntos se decidió realizar un estudio de la relación entre las instancias de forma no informada con la tarea de aprendizaje no supervisada, agrupamiento.

Otra parte de este trabajo se centra en el análisis del sentimiento de los usuarios de Twitter para obtener las impresiones de la sociedad española sobre los partidos políticos que la representan. Una vez analizados los resultados y haber clasificado las opiniones en positivas, negativas o neutras se intentaron explicar los resultados obtenidos mediante la relación con eventos ocurridos en el ámbito político.

Finalmente, respecto a la clasificación, se obtuvo un clasificador basado en *Naive Bayes* con un 98.42% de acierto al clasificar los *tweets* relacionadas con la política, economía, deportes y tecnología. Sobre el agrupamiento, no se pudieron extraer conclusiones claras dados los resultados obtenidos, los cuales son comprensibles dada la naturaleza de la tarea desarrollada. En el análisis del sentimiento se relacionaron varios resultados obtenidos de las opiniones de los usuarios de Twitter con eventos políticos ocurridos el mismo día de la recopilación de datos. Este hecho respalda que, mediante esta herramienta se pueden conocer las opiniones de los usuarios con una certeza y seguridad considerable.

Tabla de contenido

Índice de tablas

Índice de figuras

1 Introducción

1.1 Motivación

Hoy en día tenemos acceso a una gran cantidad de información gracias a la tecnología y las herramientas que se han desarrollado, las cuales ofrecen nuevas oportunidades a la sociedad, tanto personal como empresarialmente.

Antes de la aparición de las redes sociales, la mayoría de la información generada emanaba de empresas, investigaciones científicas o procesos administrativos de grandes empresas, pero esta tendencia ha cambiado. A día de hoy, la gran parte de los datos y de la información que se aporta a la sociedad proviene de las redes sociales.

Dado que un alto porcentaje de las personas que configuran la sociedad actual en la que residimos son usuarios de las redes sociales, los datos que se encuentran en estas plataformas no solo son consumidos por los usuarios, sino que también son generadores de información muy relevante. Así, todos los datos que aportan los usuarios y que se almacena en estas plataformas hacen posible que nuestros comportamientos puedan ser estudiados y empleado por muchas empresas para adaptar sus servicios a nuestras necesidades.

Nuestros comportamientos en las redes sociales, *hashtags* que empleamos, búsquedas que realizamos, los *me gustas* que realizamos a una publicación o los *retweets* crean una huella o marca digital facilitando como es el usuario, sus inquietudes, intereses, hobbies o deseos. La forma actual de analizar todos esos datos que aportamos los usuarios sin darnos cuenta está muy relacionada con el análisis del sentimiento. Mediante esta herramienta se pueden analizar las opiniones o impresiones de un usuario sobre un tema concreto, un servicio, un producto o su preferencia política, como se verá más adelante y que se analizará en este trabajo.

Por otro lado, si se quiere obtener información relevante sobre todos estos datos, éstos no sólo deben ser almacenados, sino que es necesario que sean procesados, agrupados y clasificados. Aquí entran en juego los procesos de clasificación y agrupación automática.

El concepto del *Big Data*, y esta generación masiva de datos que vivimos actualmente trae consigo nuevos métodos y técnicas para obtener un análisis óptimo, gestión y almacenamiento de los datos. Cada vez se muestra mayor interés por los datos y su descomposición para la obtención de información útil, dado que, este análisis y tratamiento es fundamental para realizar predicciones, detectar patrones o ayudar a tomar decisiones a las empresas.

Este tipo de técnicas ha crecido mucho durante los últimos años, y es un campo en crecimiento porque las empresas tanto a nivel nacional como internacional cada vez invierten más recursos en analizar los datos que generan ellos mismos y sus competidores con el objetivo de crecer y ser mejores.

1.2 Descripción del problema

En la actualidad, se generan una gran cantidad de datos cada día. El texto es uno de los formatos que más se utilizan para presentar los datos. Las redes sociales son uno de los orígenes más importantes de generación de datos, dado que permite interactuar a usuarios de cualquier parte del mundo e interconectarlos entre sí.

En 2016 se estimó que en la red social Twitter, había 328 millones de usuarios que generaban unos 65 millones de tweets [1]. A día de hoy, la cantidad de usuarios es prácticamente la misma que entonces.

El hecho de generar tal cantidad de datos, hace que la gran mayoría pasen desapercibidos y se pierda información muy valiosa ya que quedan sin clasificar, analizar, ordenar o procesar.

1.3 Objetivos

Dado que, la pérdida de información útil en las redes sociales por la falta de análisis, clasificación y procesado de los datos es un hecho, los principales objetivos del presente trabajo son los siguientes:

- Recopilación de tweets para la implementación posterior de un modelo de clasificación de confianza. Con un modelo de confianza se puede agrupar en función de la temática del texto para conseguir mejores consultas.

- Extracción de información útil e implantación de los enlaces entre los datos de la clasificación y agrupamiento.

- Análisis de los textos de los usuarios de Twitter ante las actuaciones políticas con el fin de extraer la opinión sobre los partidos políticos.

- Emplear el análisis del sentimiento para prever resultados de las votaciones como los sondeos o encuestas, es decir, mostrar que este tipo de análisis se emplee como otra herramienta válida igual que las mencionadas anteriormente.

- Relacionar los eventos políticos que se han dado en un periodo de tiempo concreto con el análisis realizado sobre la polaridad de los tweets extraídos de Twitter.

1.4 Marco Regulador

La finalidad de este proyecto es emplear diferentes técnicas de inteligencia artificial para obtener un sistema capaz de realizar clasificaciones mediante aprendizaje supervisado y no supervisado de textos extraídos de Twitter y analizar las opiniones en el ámbito político de los internautas de esta red social.

Para la realización de los diferentes estudios y experimentos que se realizan con el sistema, se emplean datos reales obtenidos de Twitter, donde además de los datos que se

analizan (principalmente, texto) hay información personal de los usuarios propietarios de dichos datos.

El problema a nivel legislativo puede achacarse principalmente a la gestión o tratamiento de la información descargada de la red social Twitter.

Por lo anteriormente mencionado es necesario que se cumpla la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal.

Con el objetivo de consolidar el cumplimiento de esta ley, se realizará lo siguiente:

- Al recoger los tweets, la información que no sea exclusivamente el propio texto de la publicación se eliminará con el fin de no manejar y guardar ningún tipo de información personal, de forma que la información gestionada por el sistema sea totalmente anónima.

- La cuenta del usuario sincronizada con el sistema para la descarga de información deberá dar su consentimiento para que dicha sincronización se lleve a cabo de forma correcta.

- Se nombra al autor del trabajo, encargado de los archivos de datos, para la manipulación responsable y seguridad de los datos.

1.5 Entorno Operacional

En esta sección se muestran los recursos empleados tanto a nivel de software como hardware para el desarrollo del trabajo de fin de grado.

La herramienta principal a nivel de hardware ha sido el ordenador del autor del trabajo que presenta las siguientes características o especificaciones:

- MacBook Pro (13-inch, 2017, Two Thunderbolt 3 ports)
    - Procesador Intel® Core i5 CPU @ 2.3GHz.
    - Memoria RAM de 8GB 2133 MHz LPDDR3.
    - Chip gráfico Intel Iris Plus Graphics 640 1536 MB
    - SO macOS Mojave 10.14.2

Las herramientas software que se han utilizado para el desarrollo del trabajo han sido:

- Microsoft Office Standard 2016:
    - Microsoft Word 2016.

    - Microsoft Excel 2016.
    - Microsoft PowerPoint 2016.
- Software para el aprendizaje automático, RapidMiner. [2]
- Complemento de análisis de sentimiento AYLIEN
- Complemento de Google Twitter Archiver
- Editor de texto Notepad

1.6 Organización del documento

En esta sección se detallará la estructura del documento explicando en qué consistirá cada una de las partes de los que está formado.

- En el primer capítulo se presenta la introducción del trabajo, la cual consta de la motivación para la elección de este trabajo, una descripción del problema que se puede resolver con la propuesta que se realiza en este documento, los objetivos del proyecto, el marco regulador y el entorno operacional.
- En el segundo capítulo se detalla el estado del arte, donde se analizan y muestran trabajos e investigaciones relacionadas con este trabajo, destacando las similitudes y diferencias con el mismo.
- En el capítulo tercero se muestra el análisis y diseño del sistema, donde se especifican las funcionalidades del sistema mediante los requisitos. Además, detalla cómo interactuará el usuario con el sistema, los casos de uso, y como está estructurado el sistema propuesto, es decir, su arquitectura.
- En el cuarto capítulo se expone de forma detallada toda la implementación, experimentación y análisis de los estudios realizados con las tareas de aprendizaje supervisado como es la clasificación de tweets de distintos temas.
- En el quinto capítulo se muestra la implementación, experimentación y análisis de los estudios realizados con las tareas de aprendizaje no supervisado como es el agrupamiento de tweets.
- En el sexto capítulo se desarrolla la implementación y experimentación del análisis del sentimiento realizado en Twitter con el ámbito político de la sociedad española de fondo.
- En el séptimo capítulo se muestran todos los datos relativos a la gestión del proyecto, donde se destacan: la planificación que se ha seguido para la realización de este, el presupuesto (con los costes desglosados en caso de ser una empresa la que desarrolla este sistema), un plan de riesgos y el impacto socioeconómico.

- El capítulo octavo se exponen las conclusiones obtenidas por la realización de este trabajo, tanto personales como técnicas haciendo referencia al proyecto, y los trabajos futuros que se podrían plantear a partir de este.

- En el noveno y último capítulo se enumeran las referencias a la información consultada para el desarrollo del trabajo.

2 Estado del arte

En este apartado se exponen en detalle los dos aspectos o temas principales que trata el presente trabajo y que están estrechamente relacionados con la minería de texto, ya que, dicha disciplina los engloba a ambos. Además, se mostrarán otras investigaciones directamente relacionadas con estos temas principales. Dichos temas son la minería de texto orientada a la clasificación y agrupación de textos extraídos de redes sociales, y, por otro lado, el análisis del sentimiento de textos extraídos de redes sociales.

Actualmente se genera una cantidad desmesurada de datos, pero la mayor parte de estos no son tratados. Los datos sin tratar no aportan ningún valor, y deben ser procesados para poder extraer información útil. Las redes sociales son un claro ejemplo de fuente de información, que, en cada instante, incrementa su volumen de datos de una forma asombrosa. Lo más importante de los datos (sociales) en sí, son las conclusiones e información que se puede extraer a partir de ellos [3].

El desarrollo de la minería de texto comienza a finales del siglo XX, cuyos sistemas o herramientas para la práctica de la misma eran costosos y primarios. Pero actualmente, dada la importancia de conocer lo que reflejan estos datos, este tipo de sistemas han evolucionado muy rápidamente junto al avance exponencial de la tecnología. En la Figura 2.1 [4] se muestra la evolución que ha tenido en los últimos años la cantidad de usuarios en redes sociales, lo que muestra la importancia de los mismos en la generación de datos.

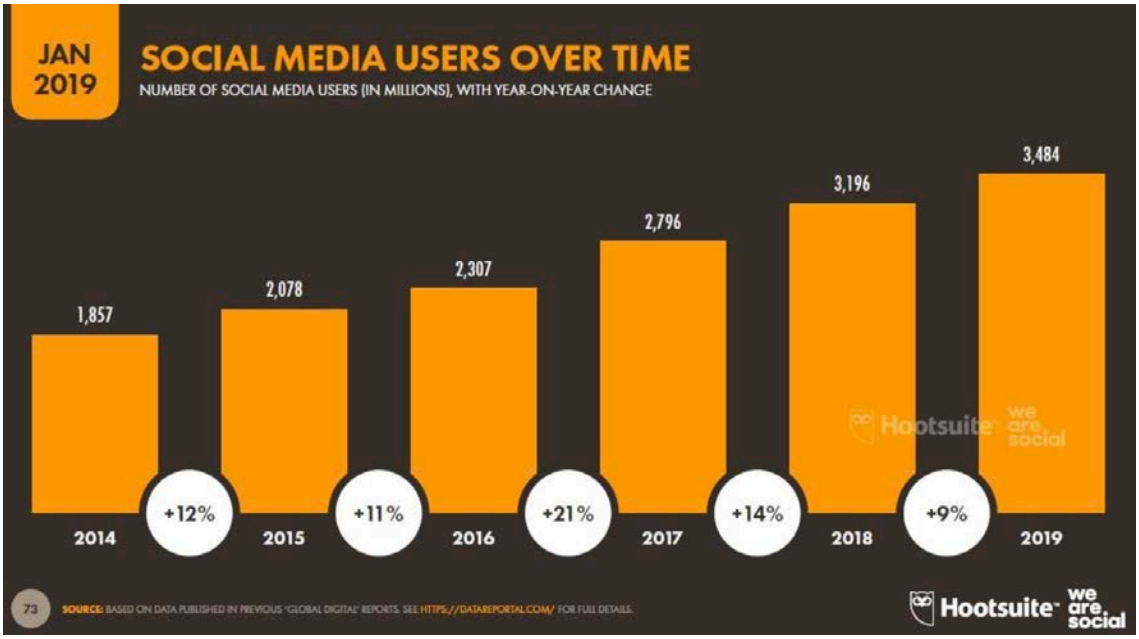


Figura 2.1 Evolución de los usuarios en redes sociales

La minería de textos se centra en extraer la información relevante y útil de textos o documentos mediante identificación de patrones en los mismos, tendencias en el uso

de palabras o el empleo de las mismas. Además, presenta ventajas importantes para diferentes ámbitos empresariales como son: [5]

- Personas que no están familiarizadas con el ámbito de la ingeniería ni del análisis de datos pueden comprender los resultados obtenidos mediante conclusiones sencillas.
- La detección de información puede ayudar a la toma de decisiones tanto estratégicas como tácticas.
- Analizando la información de los clientes, se puede mejorar el servicio prestado y de esa forma, mejorar la relación con el cliente.
- Puede ahorrar costes a la empresa y brindar nuevas oportunidades de negocio.

La clasificación automática de textos ha obtenido mucha importancia en los últimos años, siendo una de las áreas de investigación que más se ha desarrollado. Esto ha ocurrido por lo que se ha comentado anteriormente, se generan volúmenes inmensos de datos en forma de texto. Una de las fuentes más importantes de textos digitales son las redes sociales.

2.1 Minería de texto en redes sociales

Los textos engloban una gran cantidad de información, que las maquinas no pueden interpretar porque son cadenas de caracteres, por lo que es necesario aplicar algoritmos o metodologías para procesar esos textos y extraer la información útil. La minería de datos y de texto está en constante crecimiento y aún más en las redes sociales, puesto que, la gran mayoría de la información relacionada con las empresas se encuentran plasmadas en textos [6].

La minería de datos emplea diferentes campos como la inteligencia artificial, el análisis estadístico y bases de datos para extraer la información que no se puede detectar fácilmente. Así, la minería de datos es capaz de extraer relaciones, comportamientos, patrones y tendencias entre los datos analizados, lo cual, forma parte para la toma de decisiones gracias al conocimiento que aporta. La minería de datos se puede considerar una técnica puntera en el campo del análisis de datos [7].

2.2 Minería de texto para la clasificación y agrupamiento en redes sociales

En un trabajo de investigación realizado por Sixto Jansa Sanz y Enrique Ortiz Torralba [8] se plantea la clasificación de tweets en función de la temática de los mismo con el fin de poder realizar búsquedas más sencillas para el usuario, dado que la información está filtrada. Primeramente, se plantea la categorización de los tweets de forma manual, para que posteriormente se utilice una clasificación automática mediante la técnica de aprendizaje supervisado basada en el Teorema de Bayes. En este caso se plantea una

estructura en forma de árbol para la clasificación, presentando 3 categorías principales (Política, Ocio y Deporte), y dentro de ellas diferentes subcategorías.

Las conclusiones obtenidas son, que si se incrementan las categorías en las que se clasifican los tweets, se obtienen mejores resultados. La clase o categoría que obtiene los tweets no clasificados siempre tendrá un número menor de instancias, dado que, no tiene ejemplos con los que entrenarse y no se genera aprendizaje. La categoría referente al tema de Ocio, presenta peores resultados porque es un tema más general y que a mayor número de instancias de entrenamiento mejores serán los resultados.

El trabajo anteriormente explicado presenta varias similitudes con el que se propone en este TFG:

- Tienen objetivos similares, o al menos, su objetivo entraría dentro de una de las posibles aplicaciones que se le puede dar al sistema que se propone en este trabajo.
- El algoritmo clasificador que se emplea es uno de los que se van a plantear en este trabajo, *Naive Bayes*.
- Se emplean conjuntos de entrenamiento y test, de la misma forma que se realizará en este trabajo.
- La fuente de datos es la red social Twitter.
- Como conclusión, afirman que un mayor conjunto de entrenamiento puede arrojar mejores resultados. Esta afirmación se relaciona con el planteamiento de emplear varios ficheros de entrada con distintas extensiones de datos, para ver cómo evolucionaría el sistema.

Las principales diferencias con este trabajo son:

- La cantidad de instancias que se manejan, dado que, en el actual trabajo se analizan de 4000 a 40000 ejemplos y propuesto en [8], sólo procesan menos de 1000 ejemplos.
- Además, en [8], solo se utiliza un algoritmo clasificador, lo cual tiene sentido porque no se buscar obtener el mejor o compararlo con otros, como sí ocurre en el trabajo que se presenta.

En el trabajo realizado por Cristina González Rubio y Julio Villena Román [9], se realiza una clasificación de tweets relacionados con el ámbito político, teniendo en cuenta los diferentes partidos políticos del país y sus representantes. Se emplea un conjunto de instancias de entrenamiento y otro conjunto de validación del modelo de clasificación. Este trabajo persigue implementar y evaluar un clasificador automático de texto, extrayendo así información útil sobre la opinión política de los tweets de los usuarios y analizar los problemas que pueden aparecer al emplear la minería de datos al implementar

un clasificador de textos de esta tipología. Las conclusiones obtenidas exponen que los resultados obtenidos por el clasificador son considerablemente buenos, combinando entre la precisión y la cobertura del sistema un porcentaje superior al 60% en la mayoría de las clases, tal y como se muestra en la Figura 2.2 .

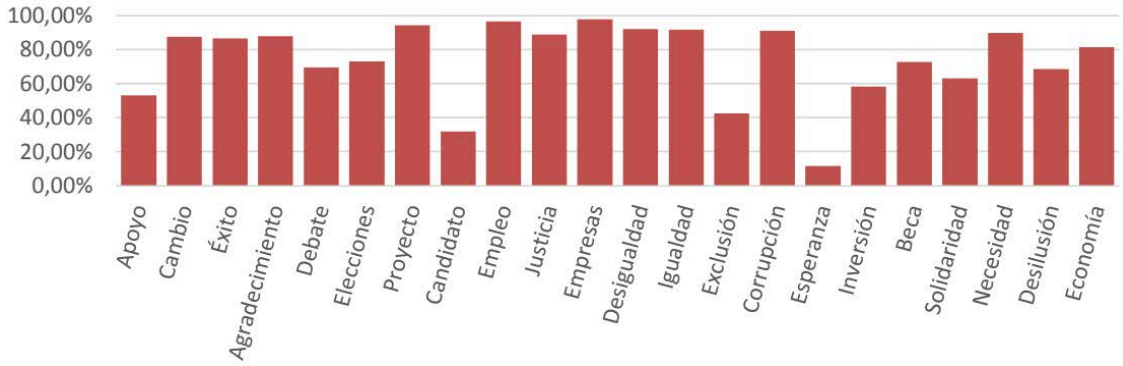


Figura 2.2 Resultados Clasificación obtenidos por Cristina González y Julio Villena

Dicho trabajo tiene varios aspectos en común con el que se presenta en este documento, los cuales se detallan a continuación.

- Los objetivos son muy semejantes, dado que, en ambos se quiere obtener un modelo de clasificación automático.
- Se emplean conjuntos de entrenamiento y test para el proceso de aprendizaje del sistema.
- El origen del que se extraen los datos es la red social Twitter.

Por otra parte, también hay diferencias entre los dos:

- La cantidad de tweets empleados está muy por debajo de la empleada en este trabajo.
- Las clases para realizar la clasificación de tweets en [9] se obtienen en el proceso de entrenamiento mediante las reglas establecidas. Esto es muy diferente al actual trabajo, dado que las clases de están definidas a priori.
- Los tweets pueden pertenecer a varias clases de forma simultánea, lo que, en nuestro caso no se posible de forma práctica.
- El modelo de clasificación es distinto a los evaluados en el presente trabajo, dado que funciona mediante reglas.

2.3 Análisis del sentimiento en textos extraídos de redes sociales.

En [10] se presenta una combinación de las dos partes que componen el trabajo presentado en este documento. Por una parte, se emplean diferentes clasificadores para realizar una clasificación sobre la polaridad de los tweets recogidos sobre el tema político electoral. Además, se emplean algoritmos de aprendizaje supervisado mediante la herramienta *WEKA*.

Las conclusiones que se han obtenido son que, los mejores modelos son las Máquinas de Soporte Vectorial (SVM) y la Regresión Logística y las Tablas de Decisión se descartan por sus malos resultados.

Este trabajo, como ya se ha denotado presenta muchas similitudes al proyecto desarrollado en el presente documento. Las principales características comunes son las siguientes:

- Análisis de la polaridad de los tweets relacionado con el panorama político.
- A pesar de ser otro apartado u otra sección del estudio, emplea algoritmos semejantes a los empleados en la sección de clasificación. Entre ellos, árboles de decisión, *Naive Bayes* y *KNN*.
- La fuente de datos es la misma red social, Twitter.

Por otro lado, también hay diferencias entre ambos trabajos:

- Los objetivos de ambos trabajos, aunque puedan parecer similares difieren entre ambos. En este [10], se categorizan los tweets con el fin de relacionar los resultados con eventos ocurridos en el clima político, mientras que, en el trabajo desarrollado en este documento, se le da mayor importancia a la clasificación y a los modelos.
- En [10] se utilizan diferentes algoritmos de clasificación supervisada, mientras que, en este, la clasificación que se realiza en este trabajo se basa en el análisis del sentimiento.
- La herramienta empleada es diferente, *WEKA*.
- La cantidad de tweets empleados es muy inferior al empleado en el presente trabajo.

Otro estudio interesante y relacionado es el presentado por Luis Deltell, Florencia Claes y José Miguel Osteso [11] en el que se estudia el empleo de la red social Twitter por parte de las principales figuras políticas durante las elecciones autonómicas andaluzas del año 2012.

En dicho proyecto se estudian los principales partidos de la comunidad autónoma en ese momento, PP, PSOE-A, IU LV-CA, UPyD, PA y eQuo, además de las cabezas visibles de cada uno de estos. Las conclusiones de dicho estudio son que mediante el análisis de los tweets y de las reacciones de los usuarios, se pueden predecir las opiniones que arrojan estos y las inclinaciones o predisposiciones políticas de la comunidad autónoma analizada.

El trabajo analizado coincide con este en los objetivos y la forma de aplicar el análisis de los tweets, ya que los relaciona con eventos ocurridos en la comunidad autónoma y lo compara con encuestas realizadas.

Las principales diferencias son: la naturaleza de las herramientas utilizadas y, por tanto, del tipo de procesamiento empleado para el análisis de los tweets. Además, tiene en cuenta diferentes campos como, los seguidores, el factor impacto o los *retweets*, los cuales no se consideran en el presente trabajo.

Para finalizar con los trabajos encontrados que guardan alguna relación con el que se presenta en este documento, se tiene una clasificación de la opinión generada en la red social Twitter respecto a temas políticos en la Ciudad de México [12]. Los autores de este proyecto desarrollaron un sistema de clasificación automática en función de la polaridad de los tweets. Recrearon un prototipo en HTML y PHP. En PHP se implementó el clasificador y en HTML la estructura. Como conclusión principal obtuvieron que mediante este tipo de análisis se pueden descubrir patrones en las formas de opinar de los usuarios dependiendo del contexto. Además, añaden que ese contexto y el tema que se obtiene de los tweets son los factores principales que condicionan la polaridad del tweet.

Este trabajo tiene un enfoque diferente al realizado, pero aporta información que se puede aplicar al análisis de la polaridad de los tweets, como son el contexto y el tema que trata.

El trabajo desarrollado en este documento presenta diferentes algunas novedades respecto a los trabajos expuestos anteriormente. Estas novedades se relacionan principalmente en la parte del análisis del sentimiento ya que en este trabajo se han relacionado los resultados obtenidos de los diferentes partidos políticos con hechos o acontecimientos ocurridos en el panorama político español. Además, en la parte de clasificación y agrupamiento se han encontrado relaciones directas entre clases, las cuales, explican dichos resultados.

3 Análisis y diseño del sistema

En este apartado se explicará el análisis y diseño del sistema. La principal función de esta sección es describir las funcionalidades del sistema y cómo serán implementadas.

El sistema deberá asegurar el cumplimiento de un conjunto de requisitos, tanto a nivel funcional como no funcional, los cuales se especificarán mediante tablas con sus características. Después se presentarán los casos de uso y finalmente la arquitectura del sistema.

3.1 Requisitos

La obtención de requisitos es una tarea imprescindible para poder definir exactamente las capacidades del sistema. Además, los requisitos son una referencia para verificar el diseño. Los requisitos se clasifican en dos categorías:

- **Funcionales**: son los que establecen la parte técnica del sistema, es decir, describen la funcionalidad del sistema.

RF-NN-VX.0	
NOMBRE	Nombre del requisito
DESCRIPCIÓN	Descripción del requisito
PRIORIDAD	□ Alta □ Media □ Baja
ESTABILIDAD	□ Si □ No
DEPENDENCIAS	

Tabla 3.1 Plantilla Requisito funcional

- **No funcionales**: describen las características y limitan como debe realizarse el desarrollo software, es decir, cómo se efectúan los requisitos funcionales.

RNF-NN-VX.0	
NOMBRE	Nombre del requisito
DESCRIPCIÓN	Descripción del requisito
PRIORIDAD	□ Alta □ Media □ Baja
ESTABILIDAD	□ Si □ No
DEPENDENCIAS	

Tabla 3.2 Plantilla Requisito no funcional

A continuación, se explican cada uno de los campos de las tablas que describen los requisitos.

- Identificador: código que identifica el requisito. El identificador cambiara en función de qué tipo de requisito sea, es decir, si es funcional será RF-NN, siendo "NN", el número del requisito, en el caso de que sea no funcional, RNF-NN. En el identificador se incorporará también la versión del requisito.

- Nombre: nombre del requisito.

- Descripción: descripción del requisito.

- Prioridad: puede tomar los valores de baja, alta y media prioridad. Se considerará alta cuando tenga mayor importancia que los demás, media, cuando no sean tan relevante como los que tienen prioridad alta, y baja cuando su importancia es mínima.

- Estabilidad: puede saberse según la vida esperada del software o puede depender de las decisiones de diseño o implementación que se vayan tomando.

- Dependencias: identificación de los requisitos con los que el requisito descrito tiene algún tipo de dependencia, indicados por su código identificador.

3.1.1 Requisitos funcionales

Los requisitos funcionales, como ya se ha explicado anteriormente, son los que hacen referencia a la funcionalidad del sistema. Con la finalidad de tener una mejor exposición y entendimiento de los mismos se presentan en forma de tabla. Los requisitos funcionales definidos son los siguientes:

RF-01-V1.0	
NOMBRE	Obtención archivo de tweets desde red social.
DESCRIPCIÓN	El sistema deberá descargar y almacenar en una hoja de cálculo los tweets obtenidos en la consulta realizada.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	---

Tabla 3.3 Requisito funcional RF-01-V1.0

RF-02-V1.0	
NOMBRE	Selección tipo de tarea.
DESCRIPCIÓN	El sistema proporcionará al usuario la opción de elegir el tipo de tarea a realizar. Estas opciones serán tarea supervisada, no supervisada o análisis del sentimiento.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	---

Tabla 3.4 Requisito funcional RF-02-V1.0

RF-03-V1.0	
NOMBRE	Selección tarea aprendizaje supervisado.
DESCRIPCIÓN	El sistema dará a elegir al usuario que tipo de proceso quiere ejecutar entre un clasificador normal o validación cruzada.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	RF-02-V1.0

Tabla 3.5 Requisito funcional RF-03-V1.0

RF-04-V1.0	
NOMBRE	Selección del clasificador en validación cruzada.
DESCRIPCIÓN	El sistema proporcionará al usuario los distintos algoritmos clasificadores disponibles para la tarea de la validación cruzada. Las opciones disponibles son *Naive Bayes, Decision Tree, KNN* y *Deep Learning*.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	RF-03-V1.0

Tabla 3.6 Requisito funcional RF-04-V1.0

RF-05-V1.0	
NOMBRE	Selección del algoritmo clasificador en clasificación general.
DESCRIPCIÓN	El sistema proporcionará al usuario los distintos algoritmos clasificadores disponibles para la tarea de clasificación. Las opciones disponibles son *Naive Bayes* y *Deep Learning*.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	RF-03-V1.0

Tabla 3.7 Requisito funcional RF-05-V1.0

RF-06-V1.0	
NOMBRE	Selección fichero conjunto de instancias.
DESCRIPCIÓN	El sistema proporcionará al usuario la opción de elegir que fichero desea emplear para los procesos disponibles. Hay cuatro ficheros con las siguientes cantidades de instancias: 4000, 12000, 20000 y 40000.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	RF-02-V1.0, RF-03-V1.0

Tabla 3.8 Requisito funcional RF-06-V1.0

RF-07-V1.0	
NOMBRE	Establecimiento palabra clave consulta Twitter.
DESCRIPCIÓN	El sistema permitirá al usuario introducir mediante un *hashtag* la clave de la búsqueda para obtener los tweets.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	---

Tabla 3.9 Requisito funcional RF-07-V1.0

RF-08-V1.0	
NOMBRE	Selección atributo a predecir.
DESCRIPCIÓN	El sistema proporcionará al usuario la posibilidad de elegir el atributo que desea predecir en las tareas de clasificación.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	RF-03-V1.0

Tabla 3.10 Requisito funcional RF-08-V1.0

RF-09-V1.0	
NOMBRE	Selección número de agrupaciones.
DESCRIPCIÓN	El sistema proporcionará al usuario la posibilidad de asignar el número de *clusters*, en los que se quieren agrupar las instancias.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	RF-02-V1.0

Tabla 3.11 Requisito funcional RF-09-V1.0

RF-10-V1.0	
NOMBRE	Selección datos de salida en tarea supervisada.
DESCRIPCIÓN	El sistema proporcionará al usuario la opción de elegir qué tipo de datos quiere obtener en la salida del proceso. Estas opciones son: matriz de confusión, *tokens* obtenidos y/o gráficas de la clasificación.
PRIORIDAD	☐ Alta ✓ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	RF-03-V1.0

Tabla 3.12 Requisito funcional RF-10-V1.0

3.1.2 Requisitos no funcionales

Los requisitos no funcionales son aquellos que limitan y describen cómo se efectúan los requisitos funcionales. Los requisitos no funcionales definidos son los siguientes:

RNF-01-V1.0	
NOMBRE	Situación estado del proceso.
DESCRIPCIÓN	Se informa al usuario del tiempo y fase por la que va el proceso con el fin de que este informado de la situación del proceso.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	---

Tabla 3.13 Requisito no funcional RNF-01-V1.0

RNF-02-V1.0	
NOMBRE	Formato ficheros de instancias.
DESCRIPCIÓN	El formato de todos los ficheros empleados para las tareas de aprendizaje supervisado y no supervisado será .xlsx.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	---

Tabla 3.14 Requisito no funcional RNF-02-V1.0

RNF-03-V1.0	
NOMBRE	Estructura ficheros para clasificación.
DESCRIPCIÓN	Los ficheros que se emplearán para la clasificación tendrán obligatoriamente como atributo la clase a la que pertenece la instancia correspondiente.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	---

Tabla 3.15 Requisito no funcional RNF-03-V1.0

RNF-04-V1.0	
NOMBRE	Estructura ficheros para agrupamiento.
DESCRIPCIÓN	Los ficheros que se emplearán para el proceso de agrupamiento no podrán contener la clase a la que pertenece la instancia correspondiente.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	---

Tabla 3.16 Requisito no funcional RNF-04-V1.0

RNF-05-V1.0	
NOMBRE	Conexión *AYLIEN*
DESCRIPCIÓN	El sistema deberá tener configurada la conexión con la herramienta externa *AYLIEN* cuando se utilice la opción de análisis del sentimiento.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	---

Tabla 3.17 Requisito no funcional RNF-05-V1.0

RNF-06-V1.0	
NOMBRE	Permisos twitter
DESCRIPCIÓN	El sistema deberá tener los permisos pertinentes de acceso a la cuenta de Twitter del usuario para poder realizar la extracción de tweets correctamente.
PRIORIDAD	✓ Alta ☐ Media ☐ Baja
ESTABILIDAD	✓ Si ☐ No
DEPENDENCIAS	---

Tabla 3.18 Requisito no funcional RNF-06-V1.0

3.2 Casos de uso

Un caso de uso consiste en la exposición de los pasos que debe realizar el actor, es decir, un usuario del sistema, para la ejecución de una tarea en el sistema.

Normalmente en los casos de uso, también se suelen diferenciar los tipos de actores/usuarios que realizan la acción, pero en este caso, solo habría un actor posible, el usuario. Por este motivo en estas tablas donde se describen los casos de uso, no se describe al tipo de usuario.

En la Figura 3.1, se muestra un esquema de los casos de uso en un nivel de detalle bajo, además se muestran los distintos casos de uso identificados y que actor le corresponde a cada uno, el usuario.

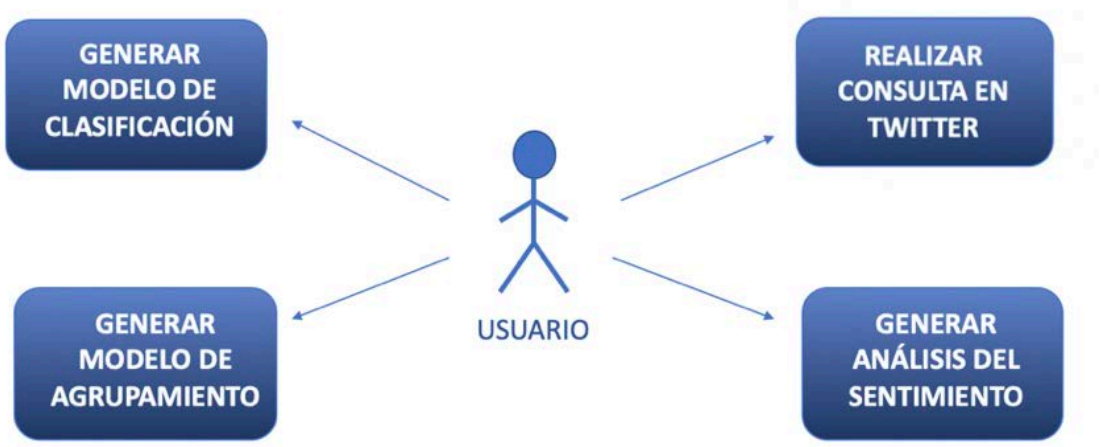


Figura 3.1 Esquema Casos de uso

CU_XX	
Caso de uso (título)	
Objetivo	
Precondiciones	
Postcondiciones	

Tabla 3.19 Plantilla Caso de uso

El significado de cada campo de la tabla es el siguiente:

- Identificador: es el parámetro que usaremos para identificar a los casos de uso, el formato que usaremos será: CU_XX, donde XX, puede ser cualquier número entero de 2 cifras.
- Caso de uso: es el nombre del caso de uso, que nos proporcionara una breve descripción del caso de uso.
- Objetivo: exposición con más detalle del caso de uso y finalidad con la que se realiza.
- Precondiciones: condiciones que deben de estar presentes para una correcta ejecución la acción.
- Postcondiciones: condiciones que se darán después de realizar la acción con éxito.
- Condiciones de fallo. Son las ejecuciones posibles que puede realizar el usuario que conllevan excepciones o fallos en el sistema.

Los casos de uso identificados son los expuestos en las siguientes tablas.

CU_01	
Caso de uso (título)	Realizar consulta a Twitter
Objetivo	Obtención de tweets filtrados por un patrón o palabra clave de la consulta establecida por el usuario para formar los ficheros de instancias.
Precondiciones	- El usuario debe tener cuenta en Twitter. - Establecimiento de permisos con la cuenta de Twitter del usuario. - Edición de parámetros de consulta.
Postcondiciones	- Obtención de un conjunto de tweets que satisfacen las claves establecidas por los parámetros de la consulta en un fichero .xlsx.
Condiciones de fallo	Parámetros incorrectos o vacíos de orden obligatorio para la ejecución de la consulta. Se obtendrá un error y no se ejecutará la consulta.

Tabla 3.20 Caso de uso CU_01

CU_02	
Caso de uso (título)	Generación de modelos de clasificación a partir de un clasificador.
Objetivo	Adquirir un modelo de clasificación empleando el aprendizaje el supervisado con la tasa de acierto de la clasificación realizada.
Precondiciones	- Selección del fichero de instancias con el que se quiere experimentar. - Elección del algoritmo clasificador con el que se quiere generar el modelo. - Modificación de parámetros del clasificador. - Ejecutar el proceso clasificador.
Postcondiciones	- Adquisición del modelo de clasificación - Adquisición de los resultados de la clasificación, tasa de acierto y matriz de confusión.
Condiciones de fallo	Parámetros incorrectos o vacíos de orden obligatorio para la ejecución de la consulta. Se obtendrá un error y no se ejecutará la consulta.

Tabla 3.21 Caso de uso CU_02

CU_03	
Caso de uso (título)	Generación de modelo de agrupamiento.
Objetivo	Adquirir un modelo de agrupamiento empleando el aprendizaje no supervisado con la distribución de las instancias en distintos conjuntos.
Precondiciones	- Selección del fichero de instancias con el que se quiere experimentar. - Edición de parámetros de *clustering*.
Postcondiciones	- Adquisición del modelo de agrupación. - Adquisición de los resultados de la agrupación como es la distribución que presentan las instancias.
Condiciones de fallo	Parámetros incorrectos o vacíos de orden obligatorio para la ejecución de la consulta. Se obtendrá un error y no se ejecutará la consulta.

Tabla 3.22 Caso de uso CU_03

CU_04	
Caso de uso (título)	Generación análisis del sentimiento.
Objetivo	Obtención del análisis de los tweets recopilados en una consulta. Los tweets se clasifican en positivos, negativos y neutros.
Precondiciones	- Vinculación del sistema con AYLIEN. - Selección de parámetros de búsqueda. - Obtención del conjunto de tweets a analizar.
Postcondiciones	- Obtención del conjunto de instancias clasificadas en positivas, negativas o neutras.
Condiciones de fallo	Parámetros incorrectos o vacíos de orden obligatorio para la ejecución de la consulta. Se obtendrá un error y no se ejecutará la consulta.

Tabla 3.23 Caso de uso CU_04

3.3 Arquitectura del sistema

En esta sección se describirá la arquitectura del sistema, explicando los distintos subsistemas o componentes por los que está conformado y cómo se relacionan entre ellos.

En la Figura 3.2 se muestra la arquitectura del sistema.

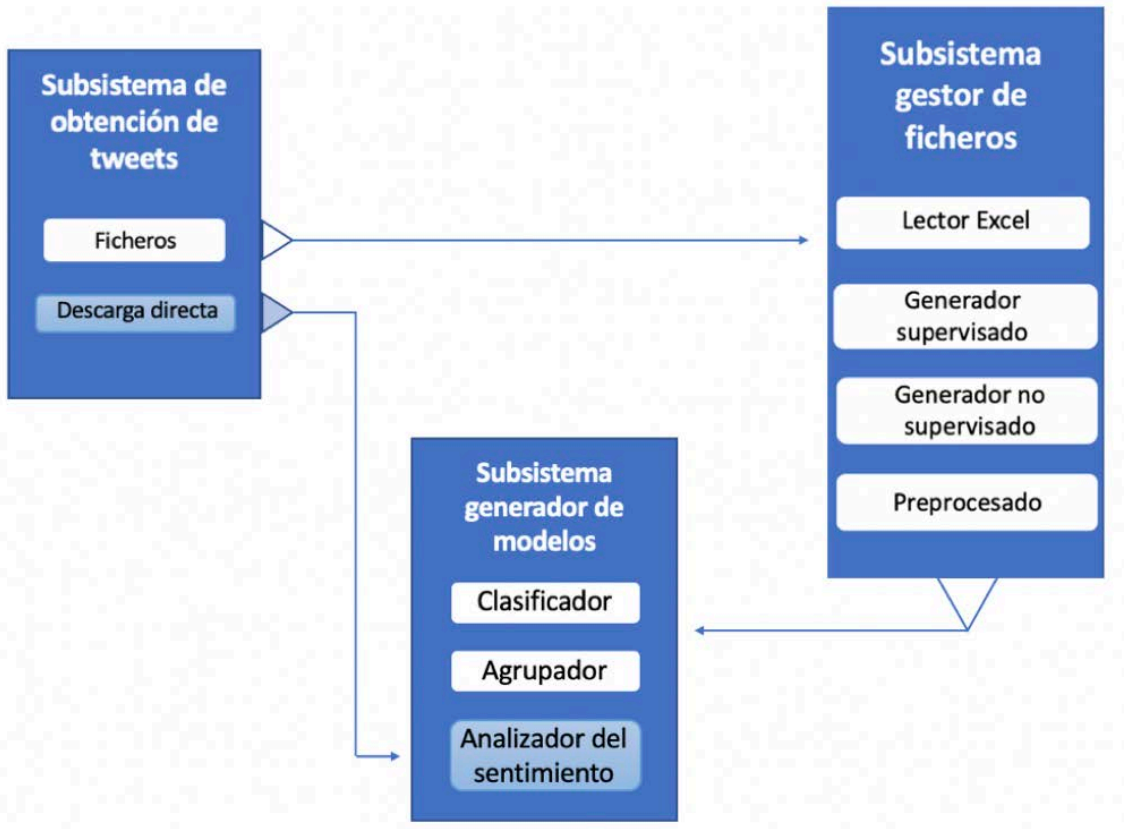


Figura 3.2 Arquitectura del sistema

A continuación, se exponen cada uno de los componentes que forman la arquitectura propuesta:

- Subsistema de obtención de tweets. Este subsistema se encarga de la tarea de recopilación de tweets de la cuenta de Twitter del usuario para su posterior utilización en las diferentes acciones disponibles. Se pueden distinguir diferentes componentes en función del destino de los tweets.
  - **ReadTwitter:** este componente se encarga de leer los tweets resultantes de la consulta.
  - **Fichero**: este componente se encarga de almacenar los tweets en una hoja con formato .xlsx para posteriormente enviarlo al subsistema gestor de ficheros.
  - **Directo:** este componente recoge los tweets de la plataforma y la deja disponible directamente en el subsistema generador de modelos para realizar exclusivamente la tarea de análisis de sentimiento.
- **Subsistema gestor de ficheros**. Este subsistema se encarga de las operaciones relacionadas con los ficheros de datos y el preprocesado de los mismos, más concretamente, con la eliminación de los hashtags de los textos. Se diferencian distintos componentes.

- **ReadExcel:** este componente se encarga de la lectura del fichero .xlsx enviado por el subsistema de obtención de tweets.

- **Generador supervisado:** este componente se encarga de darle el formato adecuado al fichero para la tarea de aprendizaje supervisado, la tarea de clasificación.

- **Generador no supervisado:** este componente se encarga de darle el formato adecuado al fichero para la tarea de aprendizaje no supervisado, la tarea de agrupamiento.

- **Preprocesado:** este componente se encarga de eliminar los hashtags de búsqueda de la consulta de los textos, una vez generado el fichero de instancias independientemente de cúal sea su tarea correspondiente.

o **Subsistema generador de modelos**. Subsistema cuya función es la de generar los modelos correspondientes a las tareas disponibles. Tiene un componente por cada una de las tareas.

- **Clasificador:** este componente se encarga de realizar las tareas de aprendizaje supervisado.

- **Agrupador**: componente cuya tarea es la generación de los modelos de *clustering*.

- **Analizador del sentimiento:** la tarea de dicho componente es realizar el análisis del sentimiento de los tweets que obtiene como entrada.

3.4 Comprensión de los datos

Los datos son parte esencial para la realización de los experimentos que se van a llevar acabo, tanto en la parte de clasificación como en las de agrupamiento y análisis del sentimiento. En este apartado se va a explicar cómo se han obtenido los datos, las herramientas utilizadas y las relaciones entre estas.

3.4.1 Clasificación y Agrupamiento

Para seleccionar los conjuntos de datos que formasen parte de esta experimentación se tomaron varios *datasets* ya generados disponibles en la red. Como el fin de la recopilación de datos es generar conjuntos de instancias fiables y reales, y éstos mostraban errores, ruido y problemas de completitud en sus instancias se optó por la recopilación propia de un nuevo *dataset*.

Además de los problemas que presentaban esos *datasets* a nivel de datos, no se adaptaban a las necesidades de la experimentación. La fórmula por la que se optó

para generar el nuevo conjunto de datos es la obtención de los datos mediante la API de Twitter. Mediante la API se podían recuperar los tweets de los usuarios filtrando por distintos campos, que, al fin y al cabo, es lo que se buscaba.

Figura 3.3 Esquema obtención de datos

1. API de Twitter

Las características de la red social, Twitter, es relevante para la obtención de los datos.

Twitter permite publicar mensajes de hasta 280 caracteres, estos mensajes aparecen en el perfil del usuario que lo publica. Por otra parte, estos mensajes se recopilan en una base de datos, la cual, presenta las siguientes peculiaridades:

- Se presentan 4 objetos principales: Publicaciones, Usuarios, Entidades y Lugares.
- La API presenta limitaciones con el fin de proteger Twitter.
- La API está basada en HTTP.
- Hay limitaciones en cuanto a la realización de solicitudes a la API.

2. Herramientas planteadas para la obtención de datos

En este apartado se presentan distintas herramientas que se han planteado para la construcción del *dataset*.

- *DataSift*: almacena todos los tweets de la red social. El inconveniente principal y que conlleva su descarte es que no trabaja con particulares, solo con empresas.
- *BackTweets*: permite buscar los últimos tweets con determinadas url´s, hashtags o usuarios. Su funcionamiento y formato los resultados de las consultas no es el adecuado.

- Programa en lenguaje de programación: es la opción que mejor se adaptaría a las necesidades dado que se diseña de cero. El problema es la complejidad que lleva hacerlo y eso se desviaría de la finalidad principal del estudio.

- *Twitter Archiver*: permite guardar y buscar tweets mediante palabras clave, hashtags o usuarios. Los tweets se almacenan en una hoja de cálculo de Google, poniendo en cada columna un dato sobre el tweet guardado.

Una vez planteadas y estudiadas las ventajas y desventajas de todas las herramientas propuestas se ha determinado que la mejor opción es *Twitter Archiver* dada su sencillez al utilizarlo y por la comodidad de descargar las búsquedas en ficheros cómodos, como son las hojas de cálculo, para la manipulación de los datos.

Habría sido interesante desarrollar un programa en algún lenguaje de programación dado que, se podrían obtener los datos con el formato que se desease y realizar los filtros que se considerasen, pero se ha estimado que tal programa no merecería la pena por el coste que tendría su desarrollo.

3. Twitter Archiver

Twitter Archiver es una extensión de Google que permite guardar consultas realizadas sobre Twitter directamente en una *Google Spreadsheet*.

Esta herramienta permite realizar distintos filtros en la búsqueda de textos como, por ejemplo, buscar por quién lo ha escrito, que ese texto este *retuiteado* o no, cuántos *retweets* tiene, por localización o por hashtags.

La herramienta presenta dos versiones disponibles, una gratuita, que permite obtener 100 tweets como máximo por hora y la opción *premium*, en la que, las consultas se realizan cada 15 minutos. La versión empleada ha sido la gratuita.

	A	B	C	D	E	F	G	H	I	J
1	Twitter Query: #deportes lang:es -filter:retweets -filter:replies									
2	Date	Screen Name	Full Name	Tweet Text	Tweet ID	Link(s)	Media	Location	Retweets	Favorites
3	11/11/2018		laFM Radio	#Deportes Ayer, el @GuadaFSF consiguió su primera victoria de la temporada. Las jerezanas vencieron 4-3 al Majadahonda.					0	0
4	11/11/2018		El Heraldo SLP	#Deportes ⚽ Confirmado: ¡hoy se juega la súperfinal Boca-River! Conoce los detalles: ⬇		http://elheraldoslp.com	https://pbs.twimg.com/		0	0
5	11/11/2018		TSiHonduras	#Deportes ⚽ Leicester empata en medio de gran homenaje a su presidente fallecido▶		http://ow.ly/RPY030mz	https://pbs.twimg.com/		0	0
6	11/11/2018		La Crónica del Quindío	#Deportes | Un empate le basta a @Cucutaoficial para asegurar su regreso a la primera división:		http://bit.ly/2DxImzN	https://pbs.twimg.com/		0	0
7	11/11/2018		Globovisión	Elvismar Rodríguez obtuvo medalla de Bronce en el Grand Prix Tashkent 2018 de Uzbekistán | #Deportes		http://goo.gl/85iFcy			0	0
8	11/11/2018		laFM Radio	#Deportes Derrota de @ascffemenino por 1-4 en su partido frente a Campo Aviación Loreto CF.					0	0
9	11/11/2018		Noticieros Hoy Mismo	#HoyMismo #Deportes ⚽ River Plate disputará su sexta final de la Libertadores, algo que solo superan cuatro equipos en la historia (Boca, Peñarol, Independiente y Olimpia); el Millonario fue campeón las últimas tres veces que llegó a la final (1986, 1996, 2015).			https://pbs.twimg.com/		0	0
10	11/11/2018		Tlatoani	#Deportes #Pumas aventaja 1-0 en el marcador ante #Toluca #TolucavsPumas			https://pbs.twimg.com/		0	0
11	11/11/2018		laFM Radio	#Deportes El @AlAndalusFemFC no pudo con el Cádiz CF Femenino y cayó por 1-3.					0	0
12	11/11/2018		Globovisión	Vargas |Más de 500 nadadores recorrieron 3 kilómetros en Naiguatá | #Deportes		https://goo.gl/eNuRZN			1	2

Figura 3.4 Formato de obtención de tweets

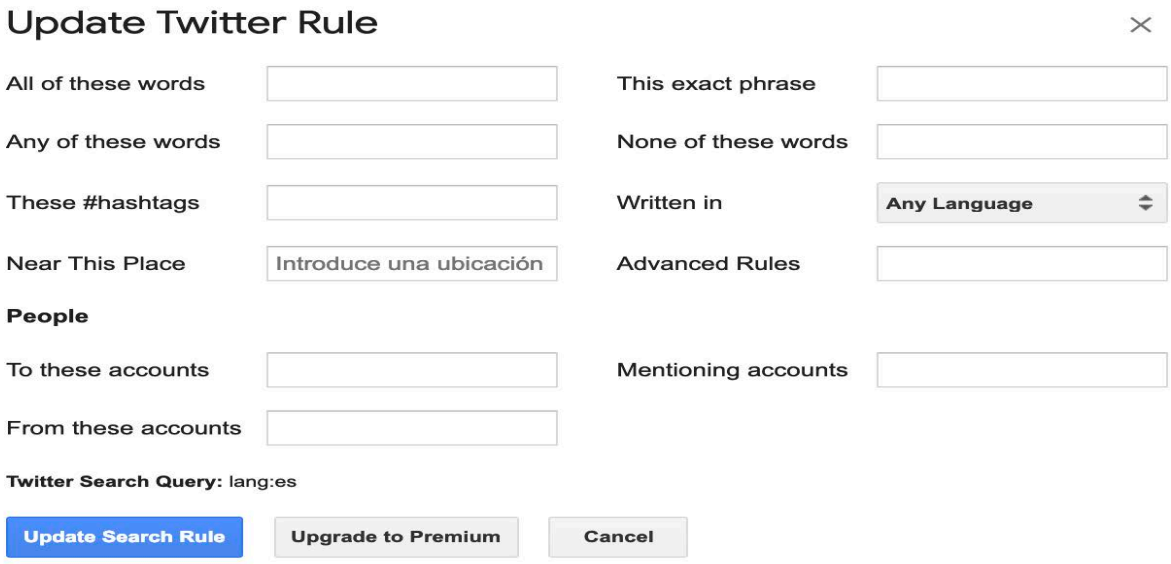


Figura 3.5 Configuración regla de búsqueda Twitter Archiver

4. Formación de los conjuntos de datos

La búsqueda de tweets se ha realizado mediante los hashtags "#Política", "#Economía"," #Tecnología" y "#Deportes", dado que son los temas sobre los que se quieren obtener, clasificar y agrupar los tweets.

El primer problema que surge es que se necesitan muchos tweets para realizar el estudio y con la versión gratuita de *Twitter Archiver* se pueden recopilar 100 tweets por hora como máximo. A esto se suma que, al realizarse mediante una cuenta de Google, solo se permite realizar una consulta por cuenta de forma

simultánea, es decir, desde una única cuenta, sólo se pueden recopilar tweets de un tema. Para que la tarea de recopilación no llevase más tiempo del deseado, se crearon 4 cuentas de Google en las que se instalaron la extensión de *Twitter Archiver*. En cada una de las cuentas se implanto la búsqueda de un tema, de esta forma, se recopilaban tweets de todos los temas de forma simultánea, acortando el coste temporal de esta tarea.

Una vez recopilado el número suficiente de tweets (realmente, los tweets no son información, serían datos de texto) se realizó un proceso en el que se eliminaron aquellos campos sobre el tweet que no interesaban para el estudio. Se estimó que el único dato importante y relevante era el propio tweet, el texto, por lo que, todos los demás campos se eliminaron.

Además, al campo del tweet se le añadió otro campo con la clase o tema al que pertenecía. De esta forma, se tenían cuatro hojas de cálculo maestras con los tweets recopilados por temáticas. Esto es importante para la formación de los ficheros con los que se experimentará posteriormente, ya que facilita una composición balanceada de las clases, la aleatorización y manipulación de los tweets.

Es preciso señalar que, en el caso de la experimentación de la clasificación y del agrupamiento no se han podido realizar algunas pares de algoritmo-conjunto. Es decir, algunos algoritmos no se han podido aplicar a algunos conjuntos de datos. Este hecho se detallará más adelante, en los algoritmos o procesos correspondientes.

3.4.2 Análisis del sentimiento

Para realizar el análisis del sentimiento de los usuarios de Twitter, se ha empleado la herramienta [2] complementada con la extensión [13]. *Rapid Miner* es una de las herramientas destinadas para la minería y análisis de datos. Mediante su entorno gráfico permite realizar diseñar o desarrollar procesos de análisis de datos. Esta desarrollado en lenguaje Java y es multiplataforma.

Por una parte, se tiene la obtención de datos mediante el acceso a datos externos de *RapidMiner*, es este caso al tratarse de Twitter, y para que se puedan descargar los tweets deseados (Figura 3.6) hay que facilitar una cuenta de dicha red social y concederle los permisos oportunos (Figura 3.7). Una vez hecho esto, se configura la conexión con la red social, la cual facilita los datos de la conexión (Figura 3.8).

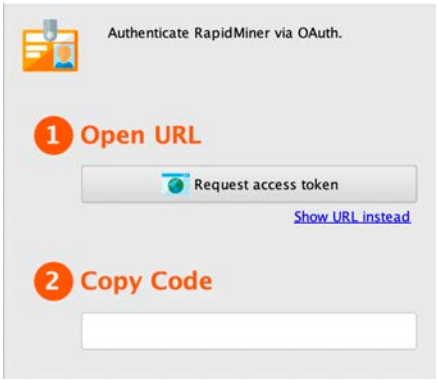


Figura 3.6 Autorización Twitter-Rapid MIner

Figura 3.7 Vinculación con Twitter y concesión de permisos

Figura 3.8 Configuración de las conexiones de AYLIEN y Twitter en Rapid Miner

Por otra parte, y con la finalidad de obtener el sentimiento de un determinado tweet, se pasa como entrada al complemento AYLIEN los tweets descargados por *RapidMiner*. Este complemento permite obtener una salida que genera el "tipo de sentimiento" que el usuario refleja con el tweet escrito. Este tipo puede tener tres posibles valores: positivo, negativo o neutro.

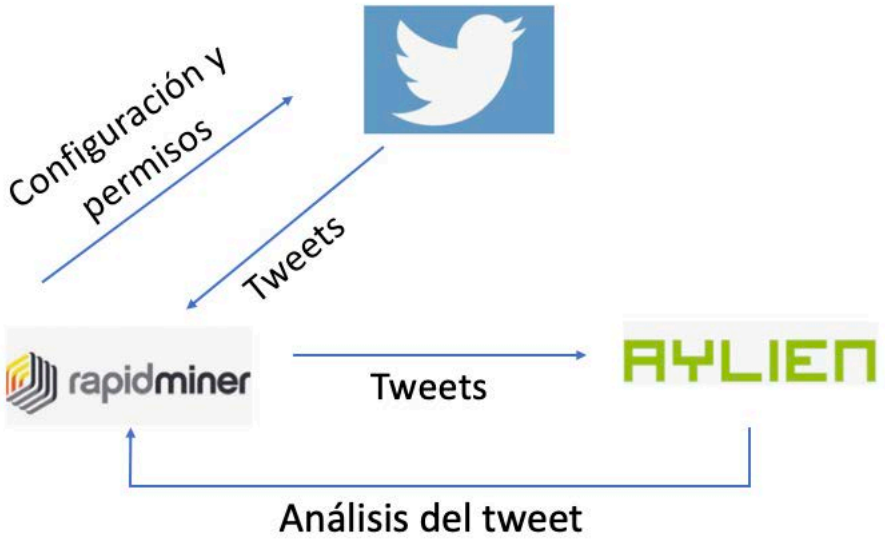


Figura 3.9 Interacción Twitter-AYLIEN-RapidMiner

De una forma muy general y simple sería como se muestra en el esquema de la Figura 3.9, pero el sistema tiene otros puntos importantes que destacamos a continuación.

RapidMiner interactúa con Twitter y con AYLIEN del siguiente modo: mediante el API de Twitter obtiene los datos, es decir los tweets de los usuarios que se quieren analizar. Para realizar esto, se utiliza el operador "*Search Twitter*", donde hay que configurar la conexión entre ambas plataformas, la cual se ha comentado anteriormente y, además, configurar las búsquedas; es decir, especificar los filtros que se quieren aplicar para obtener los tweets que se desean.

Para realizar la consulta hay que especificar en el campo "*query*" por qué palabra debe buscar los tweets. En este caso, se especificaron mediante hashtags los nombre o siglas de los partidos políticos de España. Además, se puede incluir que se descarguen los tweets re tuiteados o que no se incluyan, en este caso se ha preferido no incluirlos, y dicho caso se lleva a cabo añadiendo en la consulta "-rt".

Otro de los filtros importantes es el tipo de resultado que se quiere obtener; es decir, si se quieren descargar los tweets más recientes o los más populares. En este experimento, se consideró interesante incluir ambas opciones para que fuesen diversas las opiniones de los usuarios y se tengan en cuenta de igual forma los tweets de un personaje público, con mayor popularidad, que los de un usuario normal.

También se puede especificar tanto el límite de tweets que se quieren descargar, como máximo 100, aunque no siempre es posible descargar 100 tweets en una única consulta. Esto se debe a que, en ese momento puede no haber tantos tweets que cumplan con las restricciones indicadas. Además, hay que indicar el idioma en el que están escritos los tweets, en este trabajo utilizaremos castellano. Para seleccionar los tweets en castellano hay que especificar el código del idioma según el código ISO 639-1.

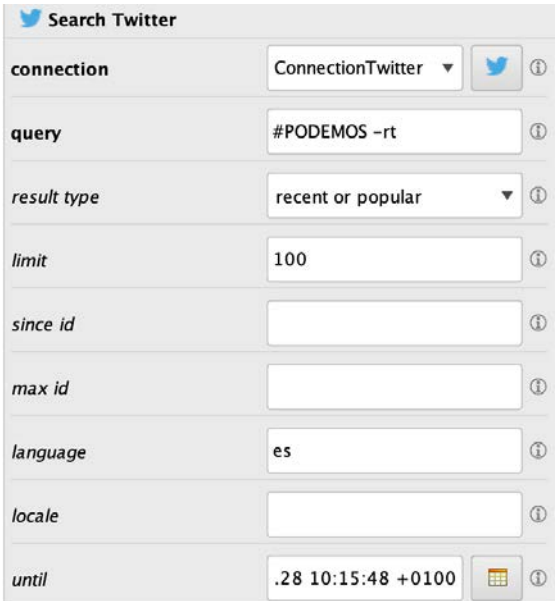


Figura 3.10 Parámetros para la formulación de la consulta a Twitter

Uno de los filtros más importantes de los que se han utilizado es el de la fecha, dado que se puede seleccionar un día concreto del que descargar tweets. Sin embargo, este filtro ha presentado un problema, ya que, si la fecha seleccionada difiere mucho de la fecha en la que se está realizando la consulta, el programa no descarga ningún tweet.

4 Implementación y evaluación: Clasificación

En la siguiente sección se explicarán y desarrollarán los procesos y estudios relacionados con la tarea de clasificación que se divide en dos partes, clasificación mediante validación cruzada y la clasificación de los mejores modelos obtenidos en la tarea anterior. Además, se hará una pequeña introducción al concepto de aprendizaje supervisado y se explicará la preparación de los ficheros empleados para dichas tareas

4.1 Preparación de datos para clasificación

Los conjuntos que se han realizado para realizar los distintos estudios o experimentos son cuatro. Cada uno de estos contiene el mismo número de instancias de cada uno de los temas o clases, están balanceadas. Cada fichero generado presenta una cantidad diferente de instancias, 4000, 12000, 20000 y 40000 instancias. Se han generado ficheros con distintos números de instancias porque se considera que la cantidad de ejemplos está relacionada con la calidad del aprendizaje del modelo generado. De esta forma se espera poder apreciar u observar una evolución en el aprendizaje de las distintas técnicas utilizadas al ir aumentando progresivamente el número de instancias.

Dado que la obtención de datos se realizó mediante la búsqueda por un hashtag, los tweets contienen en sus textos los hashtags por los que se hizo el filtro. Al lanzar los primeros procesos de la experimentación se observó que, el hecho de que los tweets tuvieran los hashtags en su contenido implicaba la clasificación automática en la clase que dicho hashtag indicaba. Este motivo fue el que hizo que se decidieran eliminar dichos *hashtags*. Este proceso se realizó de forma manual utilizando un editor de texto. Una vez preparados los tweets de todos los temas, en cada una de sus hojas de cálculo maestras se sometieron a un proceso de aleatorización con el fin de construir conjuntos de instancias que contuviesen una mayor diversidad de subtemas dentro del tema general.

Finalizando este proceso se pasó a seleccionar los tweets y crear los cuatro ficheros para realizar los estudios. Como la clasificación es una tarea de aprendizaje supervisado, además se añadió el otro campo que contienen las hojas de cálculo, la clase a la que pertenece cada tweet, la cual será la que debe predecir el sistema.

Por último, en cada uno de los ficheros se vuelve a realizar un proceso de aleatorización para mezclar los tweets y no estén ordenados por clase. Los ficheros creados son comunes para los dos procesos o estudios que se realizan dentro de la tarea de clasificación.

4.2 Aprendizaje Supervisado

El aprendizaje supervisado es un tipo de aprendizaje que se centra en revelar la relación existente entre los parámetros de entrada y unos parámetros de salida. El aprendizaje surge mostrándole a este tipo de algoritmos, cuál es la salida que se quiere para un valor concreto. Tras mostrarle una gran cantidad de ejemplos, dándose condiciones adecuadas,

el algoritmo dará un resultado correcto, incluso cuando se le muestren valores que no se haya encontrado antes.

La clave del aprendizaje es que, mediante la observación el algoritmo generaliza el conocimiento. Este aprendizaje se denomina supervisado porque al mostrarle los resultados que se desean al algoritmo se está participando en la supervisión de su aprendizaje.

4.3 Clasificación: Validación Cruzada

En el siguiente apartado se presentará una breve explicación de que es la validación cruzada y como funciona, el proceso que se ha realizado en *RapidMiner* y los algoritmos y conclusiones obtenidas en la realización de la experimentación con los mismos.

4.3.1 Introducción a la validación cruzada

La validación cruzada es un procedimiento o técnica de evaluación de un proceso analítico que pretende garantizar que los resultados no son dependientes de la distribución de los datos de prueba y de entrenamiento.

Este procedimiento consiste en obtener dos conjuntos separados partiendo de uno original, al primero se le emplea para el proceso de entrenamiento y al segundo para el proceso de validación. El conjunto de entrenamiento se divide en k subconjuntos, de forma que, se tomará cada uno de estos subconjuntos como conjunto de prueba y los demás como conjunto o muestra de entrenamiento.

El proceso se ejecuta k veces, una por cada subconjunto definido, y en cada iteración se define un conjunto diferente de pruebas. Cuando se han finalizado las k iteraciones, se calcula la precisión y el error para cada uno los modelos de clasificación y se selecciona el que presente mejor precisión y menos error promedio.

4.3.2 Implementación

A continuación, se mostrarán las distintas fases que conforman el estudio, desde que se leen o reciben los datos hasta que se obtienen los resultados de la clasificación.

Desde una perspectiva general se pueden apreciar 3 etapas:

- Entrada de los datos
- Pre procesado de los datos
- Modelización y validación cruzada

La **¡Error! No se encuentra el origen de la referencia.** Figura 4.1 muestra de forma conceptual los distintos pasos de los que se componen el proceso de clasificación

mediante validación cruzada. Se comienza por la selección del fichero, a la que le sigue su lectura. Después se realiza un pre procesado de los datos y se crea el modelo de clasificación. Finalmente se proporcionaría la salida, los resultados.

Figura 4.1 Proceso completo Validación Cruzada (Conceptual

A continuación, se describen cada uno de los pasos del proceso en *RapidMiner*, con todos los operadores empleados y sus funciones principales.

- Lectura del fichero de entrada

  Se emplea el operador "*Read Excel*", al que se establece como parámetro el fichero Excel con el conjunto de datos.
  Una vez seleccionado el fichero, se eligen o determinan los roles de cada atributo que presenta el fichero, en este caso, la clase que se quiere predecir.

Figura 4.2 Operador Read Excel

- Pre procesado de los datos

  El pre procesado está conformado por dos partes:

  - Operador "Nominal to Text": convierte los atributos nominales en atributos de cadena de caracteres. El empleo de este operador ha sido necesario para dar un formato de entrada válido al conjunto de datos de cara a los subprocesos de pre procesado. Estos subprocesos solo aceptaban como entrada atributos de tipo texto.

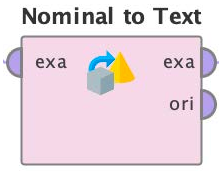


Figura 4.3 Operador Nominal to Text

  - Operador "*Process Documents from Data*": genera vectores de palabras con los atributos. Este operador tiene varios subprocesos.

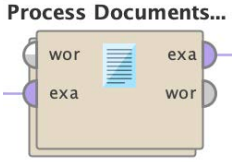


Figura 4.4 Operador Process Documents from Data

- *Tokenize*: este operador divide el texto un archivo o fichero en un conjunto de *tokens*. En este caso se ha empleado la configuración que viene por defecto, la cual separa los *tokens* en función de caracteres determinados. Se ha considerado la mejor opción porque se podrían obtener palabras clave para la clasificación.

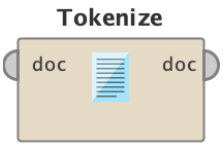


Figura 4.5 Operador Tokenize

- *Transform Cases*: transforma todos los caracteres de un documento a mayúsculas o minúsculas. De esta forma todos los *tokens* presentan la misma forma y se evitan problemas al reconocer las mayúsculas y minúsculas.

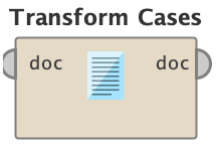


Figura 4.6 Operador Transform Cases

- *Stem (Snowball):* deriva las palabras aplicando diversos algoritmos escritos para el lenguaje o idioma seleccionado, en este caso, el castellano. Con el *Stemming* lo que se consigue es reducir los *tokens* a su raíz, lo cual puede ayudar a la recuperación de información y establecer relaciones de palabras pertenecientes a la misma familia léxica.

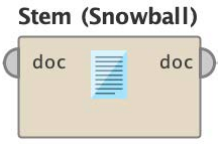


Figura 4.7 Operador Stem

- Filter Stopword: elimina del conjunto de tokens las palabras pertenecientes a una lista pasada por parámetro. Se ha empleado la lista de stopwords del castellano. Estas palabras se eliminan porque se considera que no aportan significado al texto. A la lista de palabras se han incluido todos los emoticonos disponibles, dado que, se han considerado como posible ruido para el sistema.

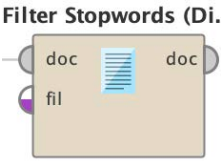


Figura 4.8 Operador Stopwords

- Operador *Cross Validation*: este operador realiza la validación cruzada para calcular el rendimiento estadístico de un modelo de aprendizaje elegido. El principal parámetro de este operador es el número de iteraciones o "*folds*" a realizar. En este caso, para la experimentación realizada se han elegido 10. Este número no se ha modificado a pesar de los distintos tamaños de los ficheros de entrada, para comparar la evolución de los algoritmos bajo las mismas condiciones. Este operador tiene dos partes:

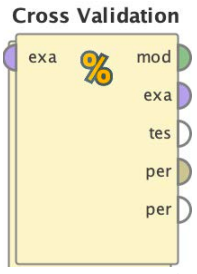


Figura 4.9 Operador Cross Validation

- *Training*: en la parte de entrenamiento se incluye el operador correspondiente al algoritmo que se quiere modelizar y el conjunto de entrenamiento. La salida es el modelo de predicción.

- *Test*: En esta parte se incluyes dos operadores:

  - *Apply Model*: tiene como entrada el modelo obtenido en el entrenamiento y un conjunto de test. Se realiza la predicción del atributo seleccionado.

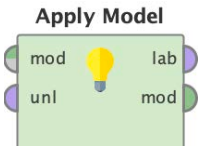


Figura 4.10 Operador Apply Model

  - *Performance*: se encarga de la evaluación del rendimiento. Ofrece los resultados de la clasificación realizada.

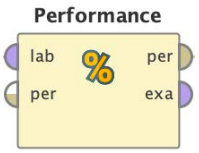


Figura 4.11 Operador Performance (Clasificación)

A continuación, se presenta el proceso completo de validación cruzada en *RapidMiner*. Se observa que el pre procesado está compuesto por un conjunto de operadores. Además, el operador correspondiente a la validación cruzada presenta las dos fases de la misma bien definidas, entrenamiento y test.

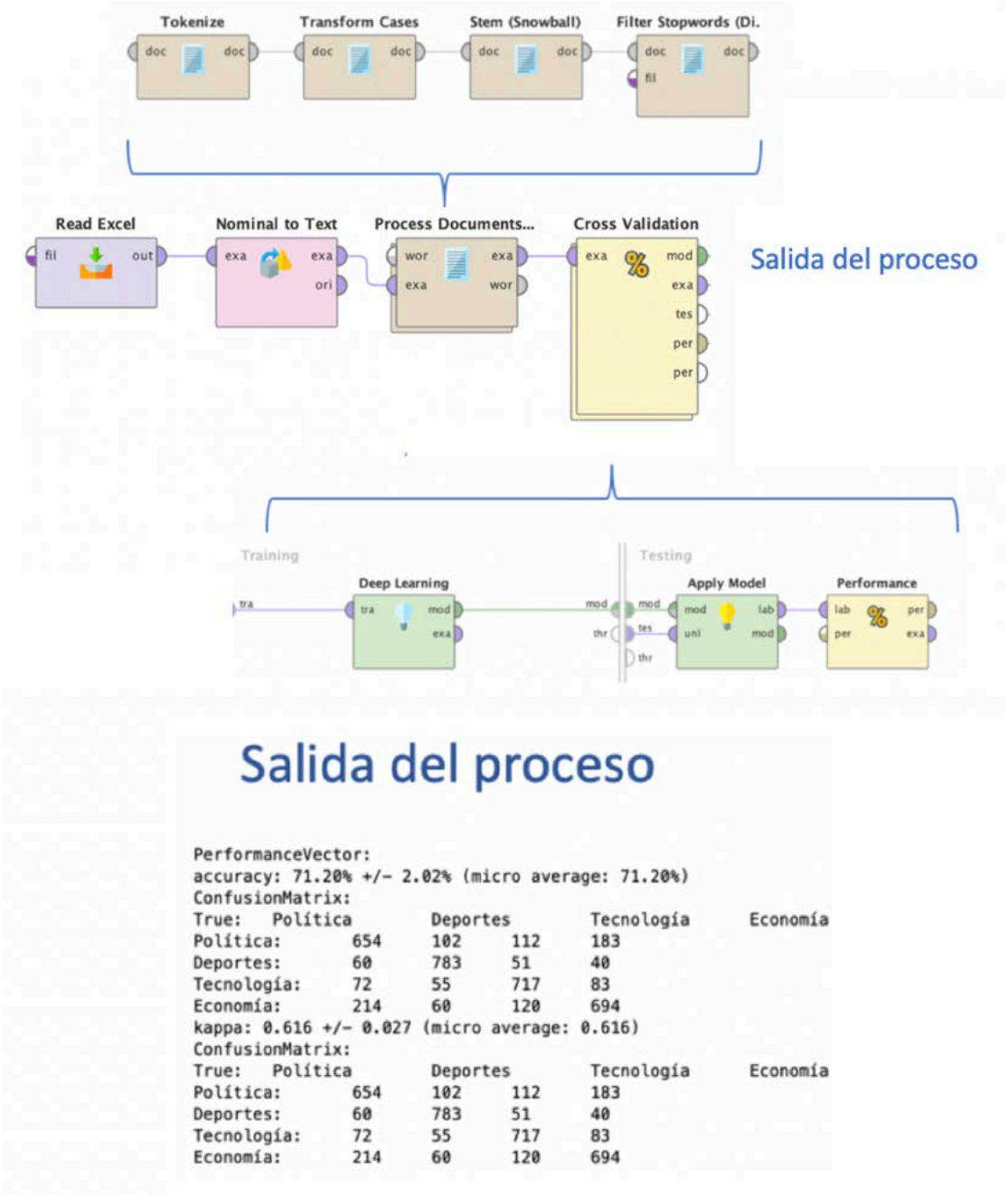


Figura 4.12 Proceso completo Validación Cruzada (RapidMiner)

4.3.3 Experimentación

En esta sección se muestran los resultados y conclusiones de los modelos obtenidos mediante la validación cruzada. Además, se realizará una breve explicación del funcionamiento de cada uno de los algoritmos seleccionados.

4.3.3.1 Naive Bayes

Este se considera uno de los clasificadores más empleados por su rapidez y facilidad.

Se define como una técnica de clasificación y predicción supervisada que construye modelos que definen la probabilidad de los posibles resultados. Está basado en el Teorema de Bayes.

El algoritmo tiene dos partes principales, por un lado, la construcción del modelo y la segunda la clasificación de un nuevo ejemplo.

Para crear el modelo se realizan los siguientes pasos:

- Calcular las probabilidades de cada clase.
- Realización de recuento de los valores de los atributos.
- Corrección de Laplace.
- Normalizar para obtener un rango de valores [0,1].

Ante un nuevo ejemplo que clasificar:

- Para cada clase, se determinan los valores de probabilidad de cada valor de los atributos del nuevo ejemplo.
- Aplicación de la fórmula de *Naive Bayes*.

A continuación, se presentan los resultados que se han obtenido en el estudio realizado con este clasificador.

Conjunto de 4000 instancias

La Tabla 4.1 Resultados Bayes-Conjunto 1 (Validación Cruzada)Tabla 4.1 muestra los resultados obtenidos por la validación cruzada con el primer conjunto de datos.

accuracy: 72.25% +/- 1.24% (micro average: 72.25%)

	true Política	true Deportes	true Tecnología	true Economía	class precision
pred. Política	676	83	93	173	65.95%
pred. Deportes	62	813	79	36	82.12%
pred. Tecnología	59	51	702	92	77.65%
pred. Economía	203	53	126	699	64.66%
class recall	67.60%	81.30%	70.20%	69.90%	

Tabla 4.1 Resultados Bayes-Conjunto 1 (Validación Cruzada)

De acuerdo a lo que se observa en la tabla se puede extraer que:

- La tasa de acierto en la clasificación es de un 72.25%. Este resultado se podría considerar relativamente positivo.

- Respecto las clases de Deportes y Tecnología, presentan los mejores resultados en la predicción de las instancias con 82.12% y 77.65%, respectivamente.

- Los peores resultados se obtienen en la predicción de los ejemplos de Política y Economía con un 65.95% y un 64.66%.

Si se comparan las instancias predichas con las correctas de cada clase, las dos clases con menos aciertos guardan relación. Entre ellas son las que más errores cometen en la clasificación. Por ejemplo, mientras que en la clase Economía, se clasifican 63 como Deportes y 126 como Tecnología, con Política se erran 203 ejemplos, un número considerablemente mayor comparado con los errores de las otras clases.

De forma inversa ocurre lo mismo, hay 173 ejemplos (tweets) son predichos como Política y que realmente pertenecen a Economía (con un número de ejemplos de 173). Sin embargo, con las otras clases se clasifican de forma incorrecta 83 y 93 ejemplos. Este hecho se vigilará en los próximos resultados de los experimentos porque de repetirse, se podría establecer una relación directa de confusión entre ambas clases y dejaría de ser un hecho aislado.

Conjunto de 12000 instancias

La Tabla 4.2 presenta los datos obtenidos con el segundo conjunto de instancias con el clasificador *Naive Bayes*.

accuracy: 80.74% +/- 1.08% (micro average: 80.74%)

	true Tecnologia	true Politica	true Deportes	true Economia	class precision
pred. Tecnologia	2385	105	108	212	84.88%
pred. Politica	121	2312	155	337	79.04%
pred. Deportes	207	167	2638	96	84.88%
pred. Economia	288	417	100	2362	74.58%

Tabla 4.2 Resultados Bayes-Conjunto 2 (Validación Cruzada)

De acuerdo a la Tabla 4.2 se observa que:

- Se obtiene un incremento un incremento considerable de la tasa de acierto promedio, obteniendo un 80.74%. Este incremento viene dado por un aumento en la clasificación particular de cada clase.

- Las clases con mejor predicción son de nuevo Deportes y Tecnología, ambas con un 84.88%.

- Política y Economía presentan valores más bajos que las otras dos clases, 79.04% y 74.58%, respectivamente.

En este caso, ocurre algo semejante a la experimentación anterior. La diferencia es que aquí se acentúa confusión en la clase Economía, al errar 417 instancias que pertenecen realmente a la clase Política. Además, comparando los fallos de las instancias clasificadas como Economía se aprecia que más de la mitad total de fallos se relacionan con Política.

En la clase Política se clasifican incorrectamente 337 instancias, que pertenecen a Economía, el doble de fallos respecto a las obtenidas con las otras dos clases.

Como se mencionó en el experimento anterior, el hecho de que las clases Economía y Política sean las que mayor confusión tengan, y esa confusión sea entre ellas, no parece una simple casualidad o un hecho aislado.

Conjunto de 20000 instancias

La Tabla 4.3 presenta los resultados que ha dado la experimentación con el tercer conjunto de datos.

accuracy: 76.21% +/- 0.73% (micro average: 76.22%)

	true Deportes	true Política	true Economía	true Tecnología	class precision
pred. Deportes	4308	322	147	221	86.19%
pred. Política	247	3189	558	253	75.09%
pred. Economía	257	1217	3928	708	64.29%
pred. Tecnología	188	272	367	3818	82.20%

Tabla 4.3 Resultados Bayes-Conjunto 3 (Validación Cruzada)

Observando la Tabla 4.3:

- El acierto promedio obtenido es del 76.21%. Este acierto desciende ligeramente respecto al estudio anterior. El descenso puede explicarse por la importante bajada en el acierto en la clase Economía, dado que los valores de las demás clases tienen valores muy similares.

- De nuevo las clases que mejores resultados arrojan son Deportes y Tecnología con valores del 86.19% y 82.20%, respectivamente.

- Política y Economía presentan nuevamente los peores resultados, 75.09% y 64.29%.

De nuevo, las clases con peores resultados son Política y Economía, aunque la mayor confusión la obtiene la segunda. Así, se predicen 1217 instancias como pertenecientes a Economía que realmente son de Política, un número de instancias considerablemente elevado comparado con los errores cometidos con las demás clases. De forma inversa

Política clasifica de forma incorrecta 517 instancias pertenecientes a Economía, menos de la mitad de errores que Economía.

Una vez estudiados los resultados, se puede concluir de nuevo que, Política y Economía presentan una confusión entre ellas, pero se puede apreciar algo más concreto. Con los presentes resultados con los vistos anteriormente, se aprecia que, el sistema clasifica muchas más instancias en Economía que realmente son de Política que de forma inversa. Este fenómeno tiene sentido y era de esperar dado que, la economía pertenece de alguna forma al entorno político y puede haber textos políticos que hablen de economía, y ese es el hecho que se refleja en el estudio.

Conjunto de 40000 instancias

La Tabla 4.4 muestra los resultados que se han obtenido con el último conjunto de instancias. Este conjunto es el que tiene mayor tamaño.

accuracy: 78.33% +/- 0.89% (micro average: 78.33%)

	true Tecnología	true Deportes	true Política	true Economía	class precision
pred. Tecnología	7724	332	377	617	85.35%
pred. Deportes	424	8831	607	256	87.28%
pred. Política	352	407	6575	924	79.62%
pred. Economía	1500	430	2441	8203	65.24%

Tabla 4.4 Resultados Bayes-Conjunto 4 (Validación Cruzada)

De acuerdo a la Tabla 4.4 se observa lo siguiente:

- Se obtiene un porcentaje de acierto del 78.33%. Presenta un pequeño incremento respecto al anterior experimento.
- Las clases con mayor acierto son Deportes y Tecnología con unos resultados que muestran un 87.28% para la primera y un 85.35% para la segunda.
- Política incrementa ligeramente su tasa particular a un 79.62% y Economía se estanca.

Se reafirma lo observado los otros experimentos realizados con Bayes, Economía y Política tienen una relación directa en cuanto a la confusión se refiere, más concretamente, Economía erra con instancias pertenecientes a Política.

4.3.3.2 KNN

KNN es un algoritmo perezoso (*lazy*), lo que conlleva que, durante el entrenamiento únicamente guarda instancias, no construye ningún modelo, no como los arboles de decisión u otros algoritmos, solo realiza la clasificación cuando llega a las instancias de test.

Al ser un algoritmo no paramétrico, el mejor modelo de datos son los propios datos. Este algoritmo tiene un aprendizaje a nivel local, es decir, la clasificación de una instancia solo depende de los K vecinos más cercanos. A diferencia de otros modelos, sobre todo de los lineales, no necesita adaptación para más de dos clases.

Los problemas que presenta KNN y que podrían explicar algunas circunstancias observadas en el estudio son su lentitud ante un conjunto grande de ejemplos de entrenamiento, conlleva una complejidad temporal y espacial importante, y es sensible a la dimensionalidad, dado que no crea un modelo global, sino que utiliza las instancias de entrenamiento para crear o representar fronteras de separación entre los valores.

Se ha considerado que estos problemas que son propios del algoritmo han podido ser los causantes de no haber podido realizar todas las pruebas que se tenían planificadas. Así únicamente se ha podido realizar la experimentación correspondiente al primer conjunto de instancias, el más pequeño. Tanto la dimensionalidad como la lentitud y complejidad afectarían directamente a las capacidades de los equipos empleados haciendo imposible el procesamiento de las pruebas.

A continuación, se mostrarán los resultados obtenidos por la experimentación aplicando el algoritmo KNN.

Conjunto de 4000 instancias

La Tabla 4.5 muestra los resultados obtenidos en la validación cruzada aplicando el algoritmo KNN al primer conjunto de datos.

accuracy: 72.10% +/- 2.16% (micro average: 72.10%)

	true Política	true Deportes	true Tecnología	true Economía	class precision
pred. Política	631	73	60	131	70.50%
pred. Deportes	81	779	58	51	80.39%
pred. Tecnología	98	75	754	98	73.56%
pred. Economía	190	73	128	720	64.81%

Tabla 4.5 Resultados KNN-Conjunto 1 (Validación Cruzada)

De acuerdo a la Tabla 4.5 se observa:

- La tasa de acierto promedio es de un 72.10%, valor considerablemente positivo.

- La mejor clasificación o mejor predicción la obtiene la clase Deportes, con un 80.39%.
- Por debajo de la primera están Tecnología y Política, con un 73.56% y un 70.50%, respectivamente.
- Los peores resultados se presentan en la clase Economía, con un 64.81% de acierto.

Se pueden obtener varias lecturas de los resultados obtenidos del proceso.

Algo relativamente llamativo es que, en los otros procesos realizados las tasas obtenidas en Deportes y Tecnología se han presentado muy positivas y similares, pero en este caso se desmarca de ese hecho y se asemeja más a la tasa de Política, es decir, la tasa de Tecnología ha bajado relativamente.

Por otro lado, se repite la relación Política-Economía que se dio en el estudio de *Naive Bayes*. Ambas clases son las que más confusión presentan, además de los errores entre ellas. La que más errores comete en la predicción es Economía con Política, es decir, se clasifican más instancias como Economía que pertenecen a Política que de forma inversa.

4.3.3.3 Deep Learning

El aprendizaje profundo o *Deep Learning* está basado en una red neuronal artificial de alimentación avanzada que contiene múltiples capas y cuyo entrenamiento se realiza mediante el descenso del gradiente de pesos estocástico utilizando propagación hacia atrás. La red puede estar compuesta por numerosas capas ocultas formadas por neuronas con diferentes funciones de activación.

Esta cascada o conjunto de capas con neuronas de procesamiento no lineal se utiliza para obtener y modificar variables. Cada capa utiliza como entrada la salida de la capa anterior. El conjunto de capas conforma una escala de características desde un grado de abstracción más bajo a uno más elevado.

En la experimentación realiza con el modelo de *Deep Learning*, debido al tipo de proceso, los conjuntos de datos y los equipos empleados, solo se obtuvieron en el estudio con el primer conjunto de instancias, con los demás, los equipos no fueron capaces de terminar los procesos.

Los resultados obtenidos de dicha experimentación se expondrán a continuación.

Conjunto de 4000 instancias

La Tabla 4.6 presenta los resultados obtenidos con la experimentación aplicando el modelo de *Deep Learning* al primer conjunto de instancias.

accuracy: 77.90% +/- 2.45% (micro average: 77.90%)

	true Política	true Deportes	true Tecnología	true Economía	class precision
pred. Política	694	55	50	139	73.99%
pred. Deportes	66	841	40	40	85.21%
pred. Tecnología	68	40	818	58	83.13%
pred. Economía	172	64	92	763	69.94%

Tabla 4.6 Resultados Deep Learning-Conjunto 1 (Validación Cruzada)

De acuerdo a la Tabla 4.6 se observa:

- El acierto promedio del modelo obtenido es de un 77.90%. Con dicho resultado se puede considerar un buen clasificador.
- Respecto a la clasificación particular en cada clase, Deportes y Tecnología, con tasas del 85.21% y 83.13%, respectivamente.
- Las dos peores clases con Política y Economía, dado que presentan tasas del 73.99% y 69.94%.

Si se observan los errores de clasificación entre clases, la clase Política y la clase Economía son las que más errores cometen entre ellas.

Este es el mismo fenómeno ocurrido en los otros dos estudios realizados con otros clasificadores. Aunque en este caso, ambas clases comenten errores semejantes, a diferencia de los otros estudios, en los que la clase que presentaba mayor confusión era Economía.

La pequeña diferencia entre los resultados de ambas clases puede deberse a que el número de instancias del conjunto de datos empleado es relativamente pequeño. Quizá con un conjunto de datos mayor la diferencia aumentaría y se obtendrían resultados semejantes a los de los otros experimentos.

4.3.3.4 Árbol de Decisión

Un árbol de decisión en una técnica o patrón de predicción que permite estudiar decisiones secuenciales basadas en el uso de resultados y probabilidades afiliadas.

Se puede ver como un conjunto de nodos cuya misión es a tomar una decisión sobre la unión de valores de una clase o una evaluación de un valor objetivo numérico.

Cada nodo representa una regla de partición para cada atributo. En la clasificación, estas reglas separan los valores que se corresponden a distintas clases. En el caso de realizar una regresión, los valores se separan para minimizar el error de manera óptima según un criterio seleccionado. En esta técnica, los nodos se crean hasta que se cumple un criterio

de parada. Las predicciones se generan en base a la etiqueta que la mayoría de ejemplos alcanzaron en la generación del nodo hoja.

Los arboles de decisión pueden procesar entradas formadas por atributos numéricos y nominales. La etiqueta debe ser nominal en el caso de la predicción o clasificación y numérica en la regresión.

En el estudio con este modelo de clasificación se han podido realizar los procesos con los tres primeros conjuntos de instancias. El hecho de no haber podido obtener resultados con todos los conjuntos de datos, se debe a los mismos motivos por los que no se pudo con *Deep Learning*, el peso computacional del proceso y las características de los equipos disponibles.

Los resultados obtenidos se presentan a continuación.

Conjunto de 4000 instancias

La Tabla 4.7 muestra los resultados obtenidos con el árbol de decisión y el primer conjunto de datos.

accuracy: 36.88% +/- 2.19% (micro average: 36.88%)

	true Política	true Deportes	true Tecnología	true Economía	class precision
pred. Política	198	168	138	188	28.61%
pred. Deportes	7	192	6	2	92.75%
pred. Tecnología	9	9	289	14	90.03%
pred. Economía	786	631	567	796	28.63%

Tabla 4.7 Resultados Decision Tree-Conjunto 1 (Validación Cruzada)

De acuerdo a la tabla 4.7 se muestra:

- El acierto promedio del árbol de decisión con el primer conjunto de 4000 instancias es de un 36.88%. El resultado obtenido se puede considerar negativo, puesto que, si fuese un 25%, se estaría ante un clasificador aleatorio. Este caso no llega a tal punto, pero no valdría como un clasificador de confianza. Si se observan los datos obtenidos, se aprecia que, a pesar de presentar esa tasa de acierto, el clasificador no es tan negativo.

- Las clases Deportes y Tecnología presentan unas tasas del 92.75% y 90.03%, respectivamente. Por lo que, en lo referido a estas dos clases, el clasificador predice correctamente casi todos los ejemplos. Estos son los porcentajes de acierto más altos de todos los modelos estudiados hasta el momento.

- Visualizando los resultados obtenidos por las otras dos clases, se aprecia el motivo de la tasa promedio obtenida. Las tasas de Política y Economía son del 28.6% y 28.63%. Por lo que cuando se trata de la predicción de las dos clases, la clasificación es prácticamente aleatoria.

  Además, aunque estas dos clases cometan errores con las otras, también se aprecia el fenómeno observado en los otros modelos, dado que, el mayor número de instancias mal clasificadas se dan entre ellas.

Conjunto de 12000 instancias

La Tabla 4.8 presenta los resultados de la experimentación realizada con el modelo de árbol de decisión y el segundo conjunto de datos.

accuracy: 33.77% +/- 0.65% (micro average: 33.77%)

	true Tecnologia	true Politica	true Deportes	true Economia	class precision
pred. Tecnologia	0	0	0	0	0.00%
pred. Politica	0	0	0	0	0.00%
pred. Deportes	12	11	1053	4	97.50%
pred. Economia	2989	2990	1948	3003	27.47%

Tabla 4.8 Resultados Decision Tree-Conjunto 2 (Validación Cruzada)

La Tabla 4.8 muestra:

- Con el conjunto formado por 12000 ejemplos, el modelo ha obtenido una tasa de acierto del 33.77%, un valor por debajo del obtenido en el anterior modelo. A pesar de incrementar el número de instancias no se ha obtenido el resultado esperado de mejorar el resultado obtenido con el primer conjunto.

- Respecto a las clases, Deportes obtiene un 97.50% de tasa de acierto, un valor muy alto, pudiéndolo clasificar casi como perfecto.

- Las otras clases presentan resultados bastante negativos, Economía tiene un 27.81% de acierto, pero los resultados no esperados son el 0% que presentan Política y Tecnología. La explicación de este hecho se podría encontrar en la parte de la clase Economía de la matriz de confusión. Tanto los ejemplos de Política como los de Tecnología se habrían clasificado como Economía.

Conjunto de 20000 instancias

La Tabla 4.9 presenta los resultados logrados con el modelo de árbol de decisión con el tercer conjunto de datos.

accuracy: 32.98% +/- 0.64% (micro average: 32.98%)

	true Deportes	true Política	true Economía	true Tecnología	class precision
pred. Deportes	1601	18	4	15	97.74%
pred. Política	0	0	0	0	0.00%
pred. Economía	3399	4982	4996	4985	27.21%
pred. Tecnología	0	0	0	0	0.00%

Tabla 4.9 Resultados Decision Tree-Conjunto 3 (Validación Cruzada)

De acuerdo con la Tabla 4.9 se observa:

- La tasa de acierto obtenida en este tercer estudio de la modelización de mediante árboles de decisión es de un 32.98%. Ligeramente más baja que en el modelo que lo precede.
- Las clases Deportes y Economía presenta valores muy similares a los del caso anterior, 97.74% y 27.21%.
- Política y Tecnología presentan un acierto del 0%. La posible explicación seria la misma que se dio anteriormente, las instancias de estas dos clases se clasifican como si perteneciesen a Economía.

  Prácticamente todos los resultados de este experimento son iguales o muy semejantes a los obtenidos con el anterior conjunto.

4.3.4 Análisis de los resultados y conclusiones generales

A continuación, la tabla Tabla 4.10 muestra las tasas de acierto promedio en función del algoritmo y el conjunto de instancias empleados. Los experimentos que no se han podido realizar por los motivos que se han comentado anteriormente se han definido con el valor "---".

	Conjunto 1	Conjunto 2	Conjunto 3	Conjunto 4
Naive Bayes	72.25%	80.74%	76.22%	78.33%
KNN	72.10%	---	---	---
Deep Learning	77.90%	---	---	---
Decision Tree	36.88%	33.77%	32.98%	---

Tabla 4.10 Resultados generales Validación Cruzada

- Relación entre las clases Economía y Política

  Dados los resultados alcanzados en cada uno de los diferentes experimentos que se han realizado, independientemente del conjunto de instancias y algoritmo

clasificador empleado, estas dos clases han presentado una relación de confusión entre sí.

Desde un punto de vista general, cultural y social, la política y la economía están estrechamente relacionas. Además, dando un paso más, la economía se podría ver como uno de los muchos campos que se ven afectados de forma directa por la política, la economía forma parte de ella a día de hoy. Se puede ver por ejemplo en el caso que se ha dado este mismo año en el que algo meramente económico, como son los presupuestos generales, ha tenido mucha más repercusión políticamente que económicamente.

Ante un tweet relacionado con algo semejante a lo que se acaba de exponer, clasificado a priori como información económica, hay muchas probabilidades de que el sistema lo tome como información política, lo cual, tendría sentido.

Esta conclusión viene precedida por el hecho de que la clasificación de ambas clases, además de la confusión mutua, ha presentado otro rasgo característico, se han cometido más errores de predicción de un tweet económico como político que de forma inversa.

- El mejor clasificador: *Naive Bayes*

El clasificador que ha presentado mejores resultados es *Naive Bayes*, alcanzando un valor máximo promedio del 80.74% en el segundo de sus experimentos.

Un dato llamativo es que, los resultados mejoran del primer estudio al segundo, algo que se esperaba al incrementar en número de ejemplos de entrenamiento. Sin embargo, ese incremento no se prolongó a lo largo de los demás experimentos. La experimentación que se esperaba con mejores resultados por el volumen de ejemplos era la correspondiente al cuarto conjunto de datos, pero no fue así, lo que no se esperaba.

Este fenómeno se puede achacar a la naturaleza de los datos. Los datos que se manejan son tweets, que al fin y al cabo son textos, los cuales presentan símbolos, palabras, letras irrelevantes o errores gramaticales, lo cual, afecta directamente al proceso de predicción dado que puede confundir al sistema.

A pesar de que, el clasificador mediante la validación cruzada no ha seguido la evolución esperada, se tomará como modelo apto para la tarea de clasificación directa.

- El peor clasificador: *Decision Tree*

  El peor clasificador ha sido el árbol de decisión, presentando valores de acierto comparables a los que presentaría un clasificador aleatorio. Se pueden destacar las predicciones de varias clases con un 0% de acierto.

  Los motivos que podrían explicar unos resultados tan negativos son que, como los conjuntos de instancias son textos y, además de lo comentado con anterioridad, pueden contener ruido a pesar del pre procesado y eliminación de este, dado que, los errores gramaticales del propio usuario o caracteres no relevantes no se han podido controlar de forma tan exhaustiva. Otro de los motivos es que el conjunto de instancias sea escaso para que el algoritmo pueda desarrollarse lo suficiente como para arrojar mejores resultados. El ultimo motivo sería que, como ha sido necesario pasar el atributo principal de nominal a texto y los arboles de decisión manejan entradas nominales haya podido crear mayor confusión o problemas al tratar los datos al sistema.

  Sería interesante poner a prueba de nuevo al clasificador con un conjunto de datos mayor, con el que su desarrollo sea optimo, y ver los resultados que se obtienen, seguramente mejores que los obtenidos.

4.4 Clasificación: Modelos seleccionados

Con los datos obtenidos mediante la validación cruzada se ha decidido ver la evolución de los mejores clasificadores.

Dado que un buen clasificador debe reducir el error del azar, se ha estimado que, para ser un clasificador apto para este proceso se debía tener al menos un estudio con una tasa de acierto superior al 75%. Al ser cuatro clases, el error del azar sería un 100% entre las cuatro clases, es decir, un 25%. De esta forma se ha establecido esa frontera de selección del 75%.

Observando los resultados de la validación cruzada, hay dos clasificadores que cumplen con la condición impuesta, *Deep Learning y Naive Bayes*.

Los ficheros o conjuntos de datos empleados en esta tarea son los mismos que se emplearon para la tarea de validación cruzada.

4.4.1 Implementación

Figura 4.13 Proceso completo Clasificación (Conceptual)

- Lectura del fichero de entrada

  Se emplea el operados "*Read Excel*", al que se le pasa como parámetro el fichero Excel con el conjunto de datos.

  Figura 4.14 Operador Read Excel

  Una vez seleccionado el fichero, se eligen o determinan los roles de cada atributo que presenta el fichero, en este caso, la clase que se quiere predecir.

- Pre procesado de los datos

  El pre procesado está conformado por dos partes:

  - Operador "*Nominal to Text*": convierte los atributos nominales en atributos de cadena de caracteres. El empleo de este operador ha sido necesario para dar un formato de entrada válido al conjunto de datos de cara a los subprocesos de pre procesado. Estos subprocesos solo aceptaban como entrada atributos de tipo texto.

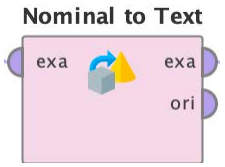
    

    Figura 4.15 Operador Nominal to Text

  - Operador "*Process Documents from Data*": genera vectores de palabras con los atributos. Este operador tiene varios subprocesos.

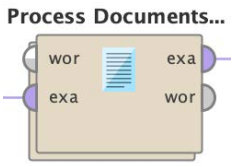
    

    Figura 4.16 Operador Process Documents from Data

    - *Tokenize*: este operador divide el texto de un archivo o fichero en un conjunto de *tokens*. En este caso se ha empleado la configuración que viene por defecto, la cual separa los *tokens* en función de caracteres determinados. Se ha considerado la mejor opción porque se podrían obtener palabras clave para la clasificación.

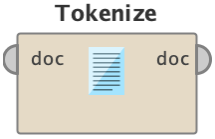
      

      Figura 4.17 Operador Tokenize

- *Transform Cases*: transforma todos los caracteres de un documento a mayúsculas o minúsculas. De esta forma todos los *tokens* presentan la misma forma y se evitan problemas al reconocer las mayúsculas y minúsculas.

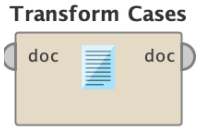


Figura 4.18 Operador Transform Cases

- *Stem (Snowball)*: deriva las palabras aplicando diversos algoritmos escritos para el lenguaje o idioma seleccionado, en este caso, el castellano. Con el *Stemming* lo que se consigue es reducir los *tokens* a su raíz, lo cual puede ayudar a la recuperación de información y establecer relaciones de palabras pertenecientes a la misma familia léxica.

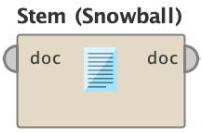


Figura 4.19 Operador Stem

- Filter *Stopword*: elimina del conjunto de *tokens* las palabras pertenecientes a una lista pasada por parámetro. Se ha empleado la lista de *stopwords* del castellano. Estas palabras se eliminan porque se considera que no aportan significado al texto. A la lista de palabras se han incluido todos los emoticonos disponibles, dado que, se han considerado como posible ruido para el sistema.

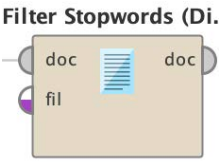


Figura 4.20 Operador Stopwords

- Modelo: El operador "modelo", se refiere al operador que hace referencia al clasificador seleccionado. En este caso será *Naive Bayes* y *Deep Learning*. Tiene como entrada la salida del pre procesado de datos y un conjunto de instancias. Las salidas son el modelo de clasificación y los ejemplos que tuvo como entrada.

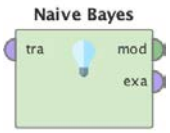


Figura 4.22 Operador Naive Bayes

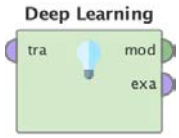


Figura 4.21 Operador Deep Learning

- *Apply Model*: tiene como entrada el modelo obtenido en el entrenamiento y un conjunto de test. Se realiza la predicción del atributo seleccionado.

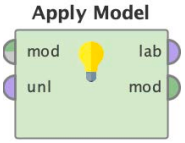


Figura 4.23 Operador Apply Model

- *Performance*: se encarga de la evaluación del rendimiento. Ofrece los resultados de la clasificación realizada.

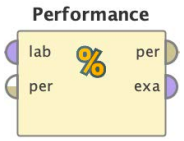


Figura 4.24 Operador Performance (Clasificación

En la Figura 4.25 se muestra el proceso completo de clasificación en *RapidMiner*.

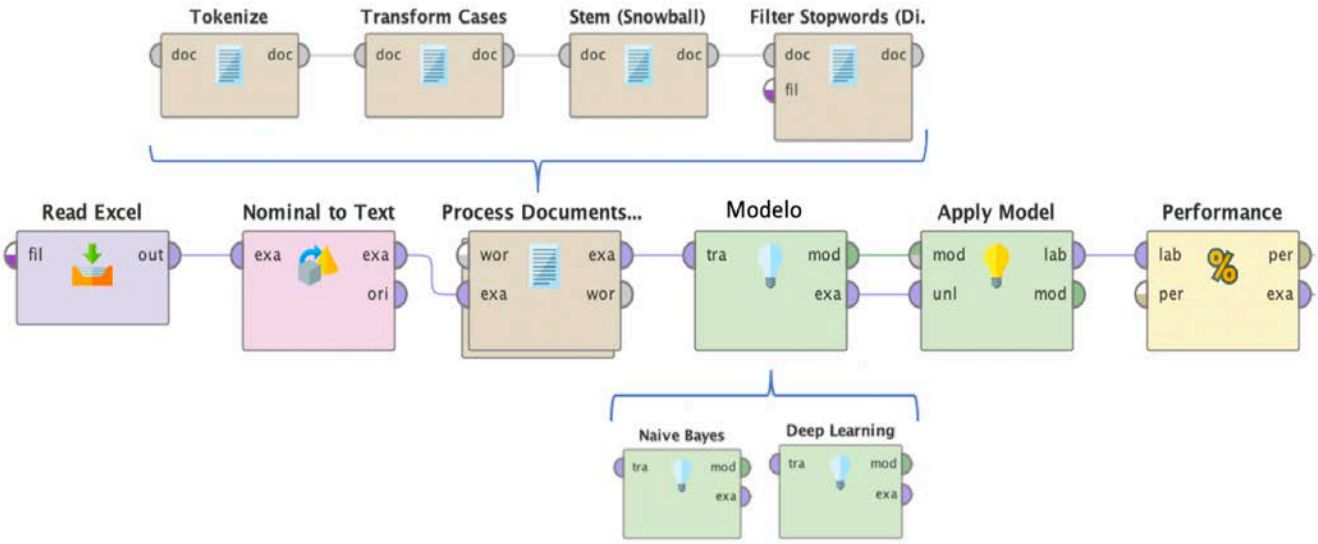


Figura 4.25 Proceso completo Clasificación (RapidMiner)

4.4.2 Experimentación

En esta sección se mostrará la evolución que han tenido los modelos o algoritmos de clasificación seleccionados por los resultados generados en la validación cruzada. Se explicarán los resultados y conclusiones tanto de los estudios con cada conjunto de datos como de manera global.

4.4.2.1 Naive Bayes

A continuación, se indican los resultados alcanzados en la experimentación de *Naive Bayes* con cada uno de los conjuntos.

Conjunto de 4000 instancias

La Tabla 4.11 muestra los resultados conseguidos aplicando el modelo de *Naive Bayes* con el primer conjunto de datos.

accuracy: 98.42%

	true Política	true Deportes	true Tecnología	true Economía	class precision
pred. Política	947	0	0	0	100.00%
pred. Deportes	0	997	4	0	99.60%
pred. Tecnología	2	0	993	0	99.80%
pred. Economía	51	3	3	1000	94.61%

Tabla 4.11 Resultados Bayes-Conjunto 1 (Clasificación)

Los aspectos más relevantes de este experimento son:

- El acierto promedio del modelo con el primer fichero presenta un 98.42%. El resultado es sorprendente y positivo, ya que, prácticamente acierta casi todas las instancias.

- Si se estudian detalladamente los resultados obtenidos en cada clase se observan fenómenos importantes. La clase Política presenta un 100% de efectividad en la predicción de sus instancias. Este hecho es sorprendente porque en la validación cruzada solía estar entre las dos peores clases.

- Las clases Deportes y Tecnología presentan una tasa de acierto del 99.60% y 99.80%, respectivamente. Ambas clases obtuvieron resultados positivos en los estudios con validación cruzada, lo que, por tanto, no es tan sorprendente.

- La clase con peor resultado es Economía, como se ha mostrado también en la mayoría de experimentos realizados. En este caso a pesar de ser la tasa con menor valor presenta una predicción muy precisa. Cabe destacar que sus errores principalmente se dan al predecir con la clase Política, lo cual no sorprende en absoluto y reafirma lo que se observó en la validación cruzada.

Conjunto de 12000 instancias

La Tabla 4.12 muestra los resultados conseguidos en la experimentación al aplicar directamente el modelo de clasificación de Naive Bayes sobre el segundo conjunto de instancias.

accuracy: 97.04%

	true Tecnologia	true Politica	true Deportes	true Economia	class precision
pred. Tecnologia	2870	12	4	15	98.93%
pred. Politica	0	2816	1	9	99.65%
pred. Deportes	86	23	2994	9	96.21%
pred. Economia	45	150	2	2974	93.79%

Tabla 4.12 Resultados Bayes-Conjunto 2 (Clasificación)

Los aspectos más relevantes que se muestran en la Tabla 4.12 son:

- Con el segundo conjunto de instancias se obtiene un acierto del 97.04%. El valor de la tasa promedio ha descendido, pero de forma muy ligera respecto al anterior experimento. Este descenso también se dio en el estudio de la validación cruzada.
- La clase con mayor tasa de acierto es Política, como en el estudio anterior, que presenta un 99.65%.
- Las clases Deportes y Tecnología obtienen un 98.18% y 98.80%, respectivamente. También se repite como el caso anterior, presentan valores muy positivos, pero muy ligeramente inferiores.
- Por último, se muestra la clase Economía que ha obtenido un 93.79% de acierto. De nuevo, casi todas las instancias erróneamente clasificadas o predichas pertenecen a la clase Política.

Conjunto de 20000 instancias

La Tabla 4.13 muestra los resultados alcanzados por el modelo generado por el clasificador *Naive Bayes* al tercer conjunto de datos.

accuracy: 95.45%

	true Deportes	true Política	true Economía	true Tecnología	class precision
pred. Deportes	4965	60	3	29	98.18%
pred. Política	2	4391	8	7	99.61%
pred. Economía	29	508	4976	206	87.01%
pred. Tecnología	4	41	13	4758	98.80%

Tabla 4.13 Resultados Bayes-Conjunto 3 (Clasificación)

Los aspectos más importantes de la Tabla 4.13 son los siguientes:

- Para el tercer conjunto de instancias, el acierto promedio del modelo es de un 95.45%. Esto se dio también en el proceso de validación cruzada,

aunque en ese caso las tasas de acierto aumentaron y descendieron. En este caso las tasas solo descienden.

- Política ha obtenido un 99.61% de acierto, que la coloca de nuevo con el mejor resultado entre las clases del experimento.
- Deportes y Tecnología presentan una tasa de acierto elevada, 98.18% y 98.80%, respectivamente.
- Economía tiene un porcentaje de acierto del 87.01%. Es la tasa más baja de todas y sigue descendiendo su valor a medida que el número de ejemplos aumenta

  Nuevamente, casi el 70% de las instancias mal predichas se relacionan con la clase Política.

Conjunto de 40000 instancias

La Tabla 4.14 indica los resultados de la experimentación realizada con el cuarto y último conjunto de datos y Naive Bayes.

accuracy: 93.51%

	true Tecnología	true Deportes	true Política	true Economía	class precision
pred. Tecnología	9249	36	69	70	98.14%
pred. Deportes	84	9848	178	23	97.19%
pred. Política	16	5	8455	56	99.10%
pred. Economía	651	111	1298	9851	82.71%

Tabla 4.14 Resultados Bayes-Conjunto 4 (Clasificación)

- Para finalizar con el algoritmo, con el ultimo conjunto de datos se ha obtenido un 93.51% de acierto.
- Política, sigue obteniendo el mayor porcentaje de acierto, como en los otros experimentos, en este caso con un 99.10%.
- Tecnología y Deportes obtienen un 98.14% y 97.19% de acierto, valores muy semejantes a los conseguidos con los otros conjuntos de datos.
- En la clase de Economía sigue descendiendo la tasa de acierto, se obtiene un 82.71%. De nuevo la mayoría de las instancias mal predichas en Economía son respecto a Política.

4.4.2.2 Deep Learning

Dada la carga computacional, la extensión del conjunto de instancias y la capacidad de los equipos empleados, no se ha podido realizar la experimentación con el último fichero.

Conjunto de 4000 instancias

La Tabla 4.15 indica los resultados conseguidos en la experimentación con *Deep Learning* y el primer conjunto de instancias.

accuracy: 98.90%

	true Política	true Deportes	true Tecnología	true Economía	class precision
pred. Política	982	0	1	8	99.09%
pred. Deportes	1	997	5	1	99.30%
pred. Tecnología	1	2	993	7	99.00%
pred. Economía	16	1	1	984	98.20%

Tabla 4.15 Resultados Deep Learning-Conjunto 1 (Clasificación)

Los aspectos más importantes que se muestran en la Tabla 4.15 son:

- La tasa de acierto promedio obtenida es del 98.90%. EL valor alcanzado es muy positivo y superior al conseguido en la validación cruzada.
- Deportes obtiene la mejor tasa de acierto, 99.30%.
- Política y Tecnología siguen de cerca al primero presentando unas tasas de acierto del 99.09% y 99.00%.
- El peor resultado lo obtuvo, como en los casos anteriores, Economía, con un 98.20%. El valor es el peor, pero es un porcentaje de acierto muy positivo. Como en otros experimentos, las instancias mal clasificadas o predichas se relacionan con Política.

Conjunto de 12000 instancias

La Tabla 4.16 muestra los resultados de la clasificación realizada con el segundo conjunto de datos.

accuracy: 87.02%

	true Tecnologia	true Politica	true Deportes	true Economia	class precision
pred. Tecnologia	2553	72	102	166	88.25%
pred. Politica	81	2699	106	268	85.57%
pred. Deportes	56	51	2679	53	94.36%
pred. Economia	311	179	114	2520	80.67%

Tabla 4.16 Resultados Deep Learning-Conjunto 2 (Clasificación)

Los aspectos más relevantes de la Tabla 4.16 son:

- En el segundo conjunto de datos se obtiene una tasa de acierto promedio del 87.02%, un valor menor que el obtenido con el primer conjunto de instancias.

- La tasa de acierto de Deportes decrece, pero no de una forma significativa, pasa a un 94.36%, además sigue siendo la clase con mayor acierto.

- Tecnología y Política si redice algo más su tasa de acierto respecto al anterior experimento, 88.25% y 85.57%.

- La tasa de Economía baja de un 98% a un 80.67%, lo que se puede considerar una reducción bastante considerable. En este caso, y de forma inesperada, la mayoría de las instancias erróneas de esta clase no se relacionan con Política sino con Tecnología. Este hecho no se había dado en ningún otro experimento, lo que se podría considerar un hecho aislado.

Conjunto de 20000 instancias

La Tabla 4.17 presentan los resultados alcanzados en la clasificación con la dupla *Deep Learning* y el tercer conjunto de instancias.

accuracy: 82.79%

	true Deportes	true Política	true Economía	true Tecnología	class precision
pred. Deportes	4217	90	49	66	95.36%
pred. Política	560	4423	862	620	68.41%
pred. Economía	68	263	3782	178	88.14%
pred. Tecnología	155	224	307	4136	85.77%

Tabla 4.17 Resultados Deep Learning-Conjunto 3 (Clasificación)

- La tasa de acierto promedio sigue con su progresivo descenso alcanzando un 82.79%. Si se compara esta tasa de acierto con la del primer experimento, se aprecia una diferencia bastante importante.

- Deportes sigue igual que en los demás experimentos, de hecho, incrementa ligeramente su tasa de acierto a 95.36%.

- Economía aumenta a un 88.14%, lo cual no se esperaba. A pesar del resultado obtenido, la mayoría de los errores que comete se relacionan con la clase Política, lo que se repite en los otros estudios.

- Tecnología desciende ligeramente a un 85.77% de acierto.

Política desciende notablemente obteniendo un valor de 68.41%. La confusión en sus instancias erróneas se relaciona con Economía, lo que era de esperar por haber ocurrido otras veces.

4.4.3 Análisis de los resultados y conclusiones generales

La Tabla 4.18 recoge las tasas de acierto de cada uno de los experimentos realizados con la clasificación en función del modelo y el conjunto empleados. Solo hay un experimento que no se ha podido realizar, y es por ello, que su valor se presenta como "---".

	Conjunto 1	Conjunto 2	Conjunto 3	Conjunto 4
Naive Bayes	98.42%	97.04%	95.45%	93.51%
Deep Learning	98.90%	87.02%	82.79%	---

Tabla 4.18 Resultados generales Clasificación

Después de haber realizado la experimentación particular con estos dos clasificadores y estudiado previamente estos y otros clasificadores en validación cruzada, se puede afirmar que las clases Política y Economía tienen una relación de confusión. Esta relación ha estado presente en casi todos los procesos realizados independientemente del clasificador y del conjunto de datos empleado. Además, en la mayoría de los resultados, las instancias predichas como Economía pertenecían realmente a Política, lo que, hace concretar más como es esta relación. Como ya se comentó, el sistema haría una interpretación en la que el tema económico podría formar parte del político, lo cual es razonable, dada la relación tan directa ambos temas.

Como se ha podido observar a medida que se ha ido incrementando el tamaño de los conjuntos de entrada, la tasa de acierto ha ido descendiendo.

- En el caso de *Naive Bayes*, estudiando las tasas particulares de cada clase se observa que, el descenso de la tasa general se debe casi exclusivamente al descenso de la tasa de Economía. Se llega a esta conclusión porque las tasas de las otras clases, a pesar de incrementar los ejemplos, casi no varían, únicamente varia de forma más notoria la de Economía. Cabe decir que, el descenso en este clasificador no es muy importante, sigue presentando resultados muy positivos.

- En el caso de *Deep Learning*, el descenso de la tasa general evoluciona diferente que en *Naive Bayes*. En este caso el descenso general si se debe a un empeoramiento de casi todas las tasas particulares. Excepto Deportes, las demás clases empeoran de un estudio a otro. Esto explicaría por qué el descenso de la tasa general es más importante, dado que de un 97.9% se pasa a un 82.79%, un 15% menos, cuando en Bayes se descendió a penas un 5%.

Ambos clasificadores serían aptos para realizar una clasificación adecuada, pero por la evolución del error y los resultados obtenidos, el mejor modelo para el sistema de predicción es *Naive Bayes*.

5 Implementación y evaluación: Agrupamiento

En el presente capitulo se explicará en que consiste la tarea de agrupamiento y los resultados obtenidos en los experimentos y pruebas realizadas.

5.1 Preparación de datos para la agrupación

Los conjuntos de datos empleados para la tarea de agrupación son los mismos que los utilizados para las tareas de clasificación.

Dado que esta tarea pertenece a las tareas de aprendizaje no supervisado, la única diferencia entre los ficheros de la clasificación y los empleados aquí es que, el atributo "CLASE" se ha eliminado.

Los ficheros, en este caso, solo se componen de un único atributo, que corresponde con el propio tweet.

5.2 Aprendizaje no supervisado

Este aprendizaje se basa en generar conocimiento únicamente a partir de los datos que se toman como entrada, sin explicarle al sistema qué resultado se quiere obtener.

Este aprendizaje busca patrones de similitud entre los datos de entrada, es decir, si una cosa es similar a otra.

5.3 K Means

El agrupamiento es una de las tareas más importantes y utilizadas del aprendizaje no supervisado.

Esta tarea se realiza mediante al método de K-medias. Dicho método comienza suponiendo que se conoce el número de grupos o clústeres en los que se tienen que agrupar las instancias. Su objetivo es encontrar la "mejor" asignación de puntos para los distintos grupos. La expresión "mejor", se refiere a maximizar las distancias inter-clústeres y minimizar las distancias intra-clústeres.

Entrando en el algoritmo, un posible método de inicialización es tomar K observaciones al azar, K definido previamente como el número de grupos. Estas observaciones se convierten en los centroides iniciales.

Para cada una de las N-K observaciones restantes, se calculan las distancias entre las observaciones correspondientes y cada uno de los centroides. Cada observación es entonces asignada al centroide más cercano. Los centroides son la observación más representativa de cada grupo.

Cuando se terminan de asignar las observaciones se tienen K grupos de observaciones.

Para cada uno de estos grupos se calcular los nuevos centroides. El proceso se repite hasta que no hay reasignaciones.

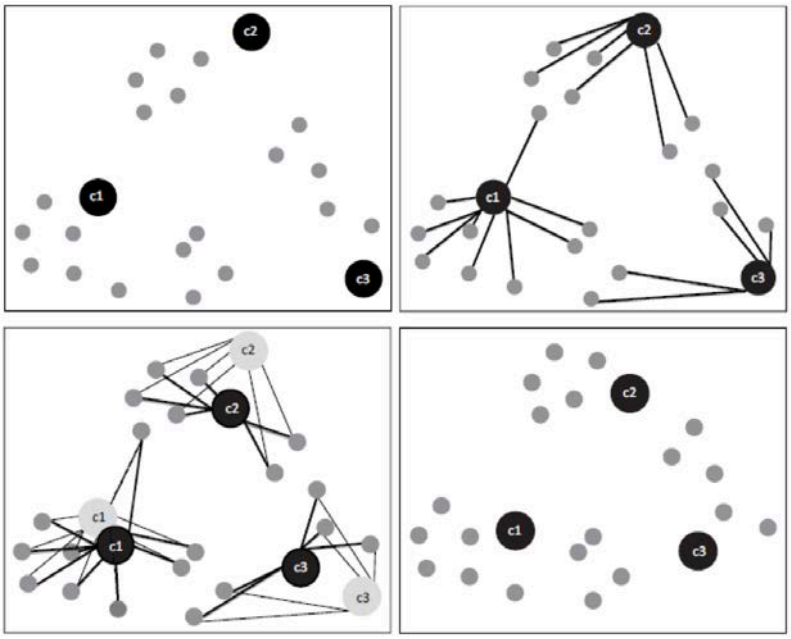


Figura 5.1 Proceso de Clustering(K-medias)

5.4 Implementación

La Figura 5.2 muestra los distintos pasos, de forma conceptual, por los que está formado el proceso de agrupamiento.

Figura 5.2 Proceso completo Clustering (Conceptual)

A continuación, se muestran las distintas etapas que sigue el proceso en *RapidMiner*, mostrando los operadores empleados y sus funciones.

- Lectura del fichero de entrada

  Se emplea el operados "*Read Excel*", al que se le pasa como parámetro el fichero Excel con el conjunto de datos.

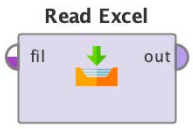
  

  Figura 5.3 Operador Read Excel

  Una vez seleccionado el fichero, se eligen o determinan los roles de cada atributo que presenta el fichero, en este caso, la clase que se quiere predecir.

- Pre procesado de los datos

  El pre procesado está conformado por dos partes:

  - Operador "*Nominal to Text*": convierte los atributos nominales en atributos de cadena de caracteres. El empleo de este operador ha sido necesario para dar un formato de entrada válido al conjunto de datos de cara a los subprocesos de pre procesado. Estos subprocesos solo aceptaban como entrada atributos de tipo texto.

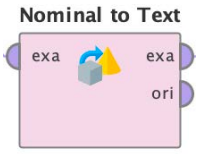
    

    Figura 5.4 Operador Nonimal to Text

  - Operador "*Process Documents from Data*": genera vectores de palabras con los atributos. Este operador tiene varios subprocesos.

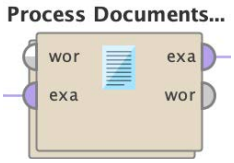
    

    Figura 5.5 Operador Process Documents from Data

    - *Tokenize*: este operador divide el texto de un archivo o fichero en un conjunto de *tokens*. En este caso se ha empleado la configuración que viene por defecto, la cual separa los *tokens* en función de caracteres determinados. Se ha considerado la mejor opción porque se podrían obtener palabras clave para la clasificación.

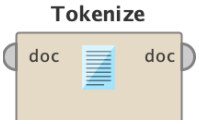
      

      Figura 5.6 Operador Tokenize

    - *Transform Cases*: transforma todos los caracteres de un documento a mayúsculas y minúsculas. De esta forma todos los *tokens* presentan la misma forma y se evitan problemas al reconocer las mayúsculas y minúsculas.

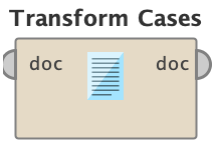
      

      Figura 5.7 Operador Transform Cases

    - *Stem (Snowball):* deriva las palabras aplicando diversos algoritmos escritos para el lenguaje o idioma seleccionado, en este caso, el castellano. Con el *Stemming* lo que se consigue es reducir los *tokens* a su raíz, lo cual puede ayudar a la recuperación de

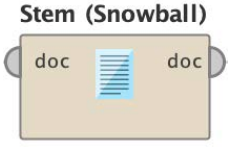
      

      Figura 5.8 Operador Stem

información y establecer relaciones de palabras pertenecientes a la misma familia léxica.

- Filter *Stopword*: elimina del conjunto de *tokens* las palabras pertenecientes a una lista pasada por parámetro. Se ha empleado la lista de *stopwords* del castellano. Estas palabras se eliminan porque se considera que no aportan significado al texto. A la lista de palabras se han incluido todos los emoticonos disponibles, dado que, se han considerado como posible ruido para el sistema.

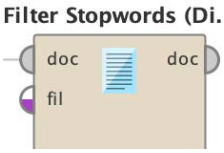


Figura 5.9 Operador Stopwords

- *Clustering*: Este operador se encarga de realizar la agrupación en función del número de clústeres que se le pasan como parámetro. Utiliza el algoritmo de K-medias. Recibe como entrada el conjunto de ejemplos a agrupar sin etiquetar.

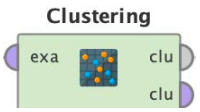


Figura 5.10 Operador Clustering

- *Performance*: Este operador toma el modelo del clúster como entrada y devuelve un vector de rendimiento que contiene los resultados obtenidos.

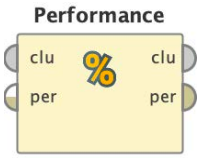
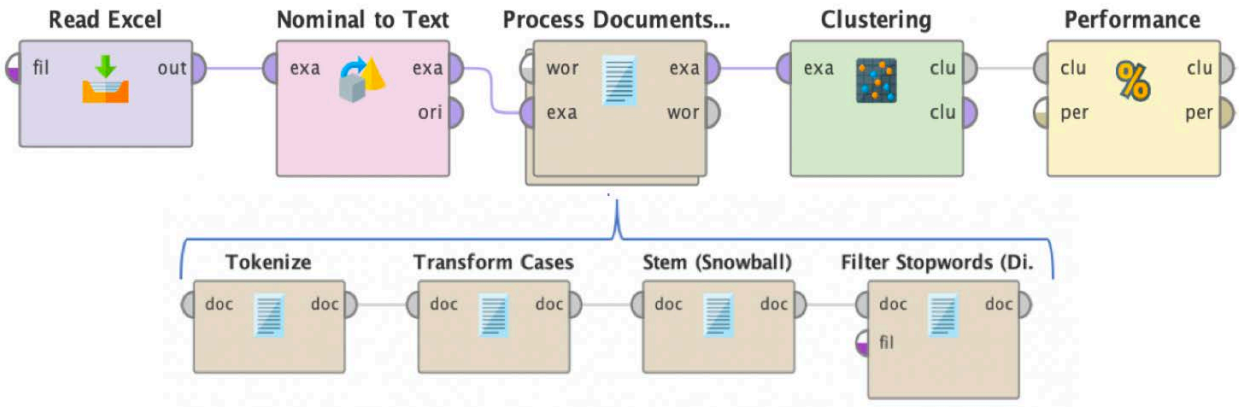


Figura 5.11 Operador Performance (Clustering)

En la **¡Error! No se encuentra el origen de la referencia.**Figura 5.12 se muestra el



proceso completo de agrupamiento en RapidMiner.

5.5 Experimentación

Dado que, las instancias pertenecen a cuatro temas diferentes, se ha determinado que el número de agrupaciones que se deben definir a priori son cuatro.

Figura 5.12 Proceso completo Clustering (RapidMiner)

A continuación, se indican los resultados alcanzados al aplicar K-medias a los distintos conjuntos de datos.

Conjunto de 4000 instancias

En la Figura 5.13 se muestra la distribución de los *clusters* alcanzada con el primer conjunto de datos.

```
Cluster 0: 3841 items
Cluster 1: 72 items
Cluster 2: 44 items
Cluster 3: 43 items
Total number of items: 4000
```

Figura 5.13 Resultados Clustering Conjunto 1

Para el primer conjunto de datos se han obtenido cuatro clústeres en los cuales se ha dado una distribución considerablemente desequilibrada, es decir, unos resultados que no nos aporta un agrupamiento claro de los diferentes tweets.

- El *cluster* 0 agrupa 3841 observaciones, lo que se traduce en que, está formado por casi todos los ejemplos del conjunto.
- El *cluster* 1 contiene 72 observaciones, una cantidad muy pequeña comparada con el tamaño del primer *cluster* y los ejemplos que debería tener.
- Los clústeres 2 y 3 han obtenido 44 y 43 ejemplos.

Los resultados, como ya se ha mencionado, no son válidos para la finalidad buscada, dado que, prácticamente todas las instancias se han agrupado de igual manera formando un grupo muy diferenciado de los otros tres.

Este fenómeno puede darse debido a que la cantidad de instancias es insuficiente para lograr esa agrupación esperada. Otro punto a tener en cuenta es la relación que se ha ido mostrando a lo largo de los estudios realizados entre las clases Política y Economía. Dado que es una tarea de aprendizaje no supervisado y la relación entre ellas el *cluster* 0 esté formado por todas las instancias de Economía y gran parte de los ejemplos de Política, además lógicamente de muchas de las instancias de las otras dos clases.

Probablemente el *cluster* 1 esté relacionado con Deportes, y el *cluster* 2 y 3 con Política y Tecnología, porque por estudios anteriores Deportes siempre se ha mantenido muy bien diferenciado de las demás clases, y Política y Tecnología es con las que más confusión se ha tenido al predecir respecto a la clase Economía.

Conjunto de 12000 instancias

En la Figura 5.14 se indica la distribución obtenida por el proceso de agrupamiento con el segundo conjunto de datos.

```
Cluster 0: 11414 items
Cluster 1: 114 items
Cluster 2: 365 items
Cluster 3: 117 items
Total number of items: 12010
```

Figura 5.14 Resultados Clustering Conjunto 2

A pesar de incrementar el tamaño de la muestra, se han obtenido resultados muy similares a los arrojados por el primer conjunto.

- El *cluster* 0 agrupa 11414 ejemplos, como en el caso anterior, agrupa prácticamente casi todos los ejemplos del conjunto.
- El *cluster* 2 está formado por 365 ejemplos. Este *cluster* presenta más del doble de ejemplos que los clústeres 1 y 3.
- Los clústeres 1 y 3 han obtenido 114 y 117 ejemplos, respectivamente.
- Si se comparan los grupos obtenidos con los del primer conjunto, los clústeres 0 coinciden, el cluster 1 del primero conjunto, en este sería el cluster 2, y los clústeres 2 y 3 del primero serían el 1 y 3 del segundo.

Al aumentar la muestra, se esperaba alguna mejora, por ligera que fuese, pero no ha sido así. El sistema no está cumpliendo con las expectativas. Para el próximo conjunto, al incrementarse de manera considerable respecto al primer conjunto, se espera alguna mejora.

Conjunto de 20000 instancias

La Figura 5.15 presenta la agrupación obtenida con el tercer conjunto de datos.

```
Cluster 0: 18993 items
Cluster 1: 333 items
Cluster 2: 471 items
Cluster 3: 203 items
Total number of items: 20000
```

Figura 5.15 Resultados Clustering Conjunto 3

De nuevo, aunque la muestra o el tamaño del conjunto de ejemplos es considerablemente mayor que la muestra empleada en el primer estudio, no se ha conseguido ningún tipo de mejora.

En los tres estudios se han obtenido resultados muy similares pero proporcionales a los tamaños de sus muestras correspondientes.

- El *cluster* 0 agrupa 184993 ejemplos, como en los casos anteriores, agrupa prácticamente casi todos los ejemplos del conjunto.
- El *cluster* 2 está formado por 471 ejemplos.
- Los clústeres 1 y 3 han obtenido 333 y 203 ejemplos, respectivamente.
- Respecto al estudio anterior se podría decir que los clústeres han cambiado ligeramente en el 1 y 3, pero siguen siendo muy similares.

5.6 Análisis de resultados y conclusiones

Los resultados obtenidos en la tarea de agrupamiento no han sido tan positivos como se esperaban. Las expectativas eran altas dados los resultados que se habían conseguido en la tarea de clasificación.

Estos resultados dejan claro que independientemente del fichero de entrada que tenga como parámetro, el agrupamiento resultante es prácticamente el mismo. Se esperaba que, al incrementar el número de instancias de un conjunto a otro, el comportamiento del agrupamiento mejorase de un experimento a otro. Este hecho no termina de sorprender, dado que, en la tarea de clasificación de esperaba también una mejoría entre conjuntos, pero se dio el efecto contrario.

Con lo expuesto anteriormente se puede concluir que el comportamiento del agrupamiento, en este caso, es independiente de los ficheros de entrada.

Además, sabiendo de antemano que dos clases presentan una relación que confunde al sistema, se podría concluir que, el hecho de que haya un *cluster* que acoge a casi todas las instancias del conjunto tendría relación con este hecho.

Por otro lado, y para finalizar, la tarea de agrupamiento al ser una tarea de aprendizaje no supervisado, es lógico que los resultados no sean tan positivos como se podían esperar.

6 Implementación y experimentación: Análisis de sentimiento

6.1 INTRODUCCIÓN

El objetivo del siguiente experimento será obtener información sobre la opinión de los usuarios de la red social Twitter sobre los partidos políticos de nuestro país. Esta información obtenida será analizada con detalle. Este experimento se ha realizado en un periodo concreto de tiempo, pero realmente lo que se desea es que, se pueda aplicar a cualquier otro momento.

La motivación de este experimento nace de la necesidad de encontrar una explicación al "comportamiento electoral" de los ciudadanos en el mes de diciembre de 2018 en Andalucía [14].

Como es conocido, estas elecciones causaron mucho revuelo porque, en general, no se supo extraer de forma anticipada información certera sobre el pronóstico de voto de los ciudadanos andaluces.

Como resultado de este experimento, se mostrará la relación de las diferentes opiniones de los usuarios de Twitter con los acontecimientos políticos que se dan en la sociedad. Se demostrará así que este tipo de análisis puede presentar resultados concluyentes de cara a futuras interpretaciones de la intención de voto, tal y como se realiza actualmente con las encuestas o sondeos. Además, es interesante destacar que el número de tweets analizados es fundamental en el resultado obtenido. Sin embargo, el proceso realizado con mayor cantidad de tweets (millones), sería similar.

6.2 Implementación

En la Figura 6.1 se presenta el esquema del proceso donde "*Search Twitter*" descarga los tweets y los analiza AYLIEN mediante el operador "*Analyze Sentiment*", cuya salida está conectada a *RapidMiner*, la cual se muestra en la interfaz de éste.

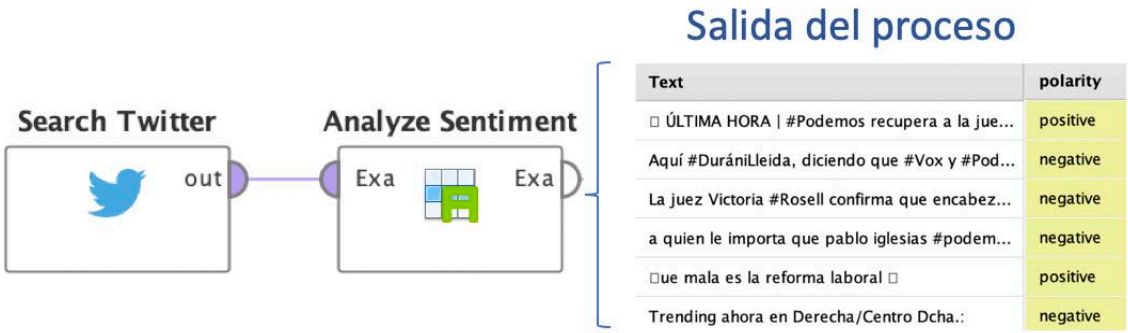


Figura 6.1 Proceso de Análisis del Sentimiento con la salida correspondiente

Hasta el momento el proceso se ha centrado en *RapidMiner*, pero realmente, el análisis de sentimiento es realizado por la extensión AYLIEN. Para ello, primeramente, se extraen las impresiones del texto que se le pasa como parámetro, en este caso tweets, y proporciona si esa perspectiva o impresión del autor es positiva, negativa o neutra. Además, se indica si el texto es subjetivo, es decir, una simple opinión o es objetivo, expresa un hecho.

También se debe destacar que AYLIEN tiene varios parámetros, tales como la conexión con *RapidMiner*, el tipo de atributo de entrada, el cual es un texto, y el modo, en el cual se puede indicar si es un tweet o un documento.

6.3 EXPERIMENTACIÓN

La experimentación ha consistido en la obtención de las impresiones (plasmada en los tweets) de los usuarios de Twitter a partir de las elecciones que se dieron en Andalucía el pasado 2 de diciembre de 2018. Estas impresiones se han obtenido mediante el proceso que se ha explicado en el apartado anterior.

Dicho proceso se ha realizado para los partidos más importantes o principales que salieron en dichas elecciones, los cuales fueron: Vox, Ciudadanos, Partido Popular (PP), Partido Socialista Obrero Español (PSOE) y Podemos. Como se ha mencionado anteriormente, este proceso se podría aplicar en otro momento y con otros partidos para obtener otras impresiones u otros análisis de distintas situaciones.

Los datos se han ido obteniendo a lo largo de los meses de diciembre y enero de 2018, de forma aleatoria y según permitía la propia herramienta. Se debe también destacar que la herramienta ha presentado varios problemas, tales como que, a partir de un número de ejecuciones, no era capaz de obtener más tweets o incluso de no especificar si eran positivos, negativos o neutros. Sin embargo, este tipo de problemas han podido solventarse obteniendo los resultados deseados.

Las consultas mediante la herramienta empleada han presentado distintos filtros, con el fin de conseguir los tweets que más interesaban o adaptaban al experimento. Para cada partido se han realizado 15 consultas en días aleatorios, como se ha comentado

anteriormente. Los filtros aplicados son, principalmente, tres: 1) búsqueda mediante los nombres de los partidos con los hashtags, 2) el idioma del tweet debe ser el castellano y, por último, 3) la fecha de los tweets, que se iba seleccionando en cada momento.

A continuación, se muestran las gráficas de cada partido con la progresión en el tiempo de los tweets evaluados como positivos, negativos y neutros, además de algunos datos obtenidos de las propias gráficas.

6.3.1 Partido VOX

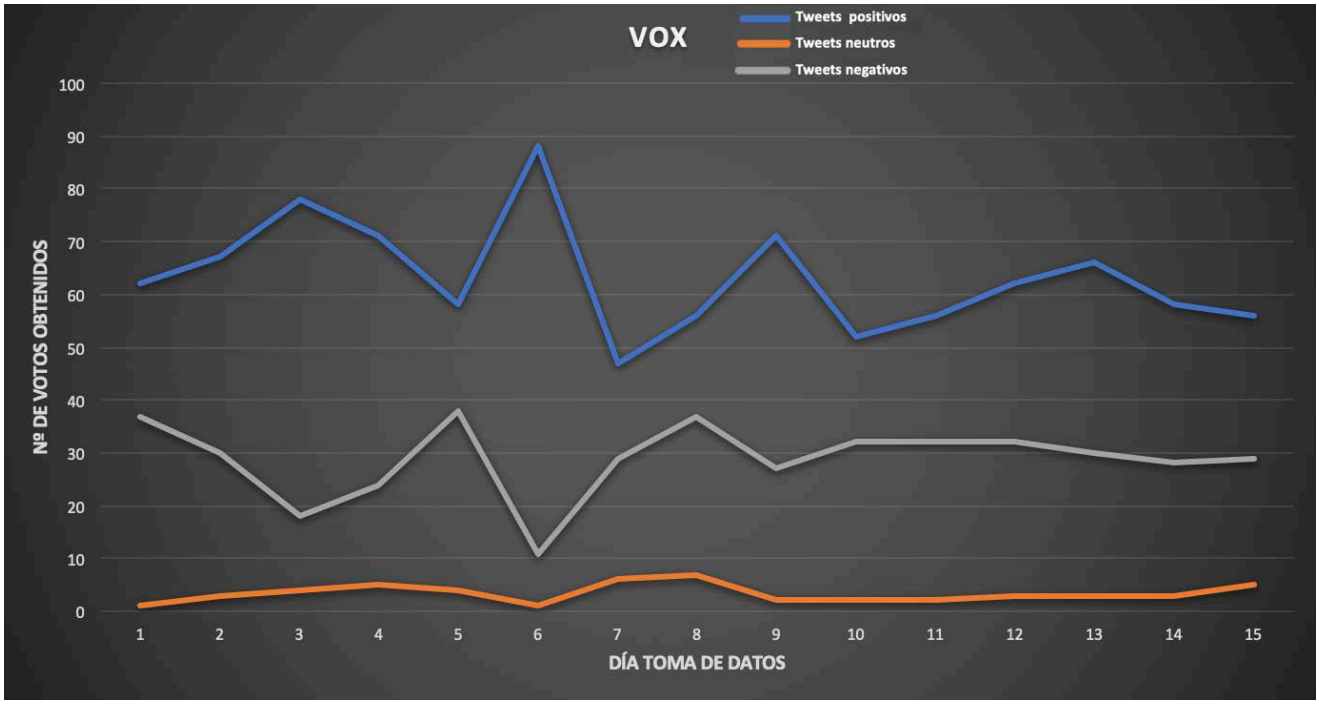


Figura 6.2 Evolución temporal de los tweets VOX

Tweets totales	1433
Porcentaje de tweets positivos	948 0,661549197
Porcentaje de tweets neutros	51 0,035589672
Porcentaje de tweets negativos	434 0,30286113

Tabla 6.1 Tabla resultados VOX

Toma de datos	Día correspondiente
1	30-12-2018
2	31-12-2018
3	01-01-2019
4	02-01-2019
5	03-01-2019
6	04-01-2019
7	22-01-2019
8	23-01-2019
9	24-01-2019
10	25-01-2019
11	26-01-2019
12	27-01-2019
13	28-01-2019
14	29-01-2019
15	30-01-2019

Tabla 6.2 Correspondencia toma de datos VOX

Para esta experimentación, se analizaron un total de 1433 tweets obtenidos a lo largo de un periodo de varios meses seleccionando quince días distintos, con el fin de obtener mayor variedad de opiniones.

A partir de estos resultados, analizaremos los resultados en función de cada uno de los partidos políticos. Así, se obtiene que en las impresiones de los usuarios sobre VOX, se presentan del siguiente modo:

- El 66,15% están categorizadas como impresiones positivas, lo que se traduce en 948 tweets
- El 3,5% clasificados son impresiones neutras, lo que supone 51 tweets
- Como negativos se han obtenido 434 tweets, un 30,35%.

Esta forma de analizar los resultados se aplicó también en la obtención de información de los demás partidos políticos.

Por otro lado, en la gráfica se aprecian distintos picos, tanto en la línea representada en color azul, las impresiones positivas como la representada en color gris, impresiones negativas. A pesar de que en ambas se presentan picos o máximos en los valores que toman, las más destacables se presentan en la línea azul, las impresiones positivas.

Un fenómeno que arroja sentido y credibilidad a los valores representados en la gráfica, es que, los picos positivos coinciden con los valores negativos más bajos y viceversa. Además de aportar sentido, se muestra que, en los puntos 3 y 6 del eje X, que representan los días en los que se han tomado los datos, las opiniones sobre VOX se presentan de forma clara como positivas.

6.3.2 Partido Ciudadanos

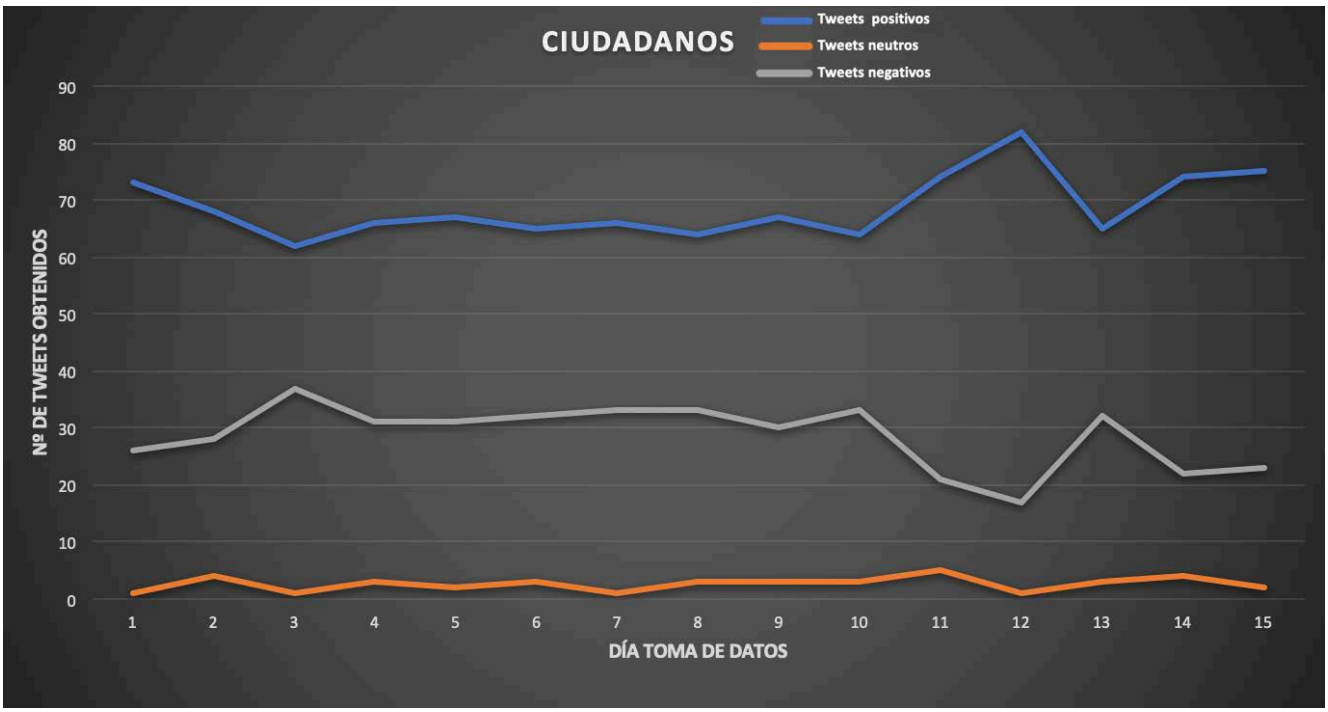


Figura 6.3 Evolución temporal de los tweets Ciudadanos

Tweets totales	1500
Porcentaje de tweets positivos	1032 0,688
Porcentaje de tweets neutros	39 0,026
Porcentaje de tweets negativos	429 0,286

Tabla 6.3 Tabla resultados Ciudadanos

Toma de datos	Día correspondiente
1	06-01-2019
2	07-01-2019
3	08-01-2019
4	09-01-2019
5	10-01-2019
6	11-01-2019
7	12-01-2019
8	13-01-2019
9	14-01-2019
10	15-01-2019
11	27-01-2019
12	28-01-2019
13	29-01-2019
14	30-01-2019
15	31-01-2019

Tabla 6.4 Correspondencia toma de datos Ciudadanos

Para el partido político que encabeza Albert Rivera se analizaron un total de 1500 tweets, los cuales se han distribuido de la siguiente forma:

- Se obtuvieron 1032 como impresiones positivas, lo que se traduce en un 68,8% del total.
- El sistema identificó 39 tweets como impresiones neutras, es decir, un 2,6%.
- Los 429 tweets restantes se consideraron como impresiones negativas, lo que equivale a un 28,6%.

Visualizando la gráfica, se aprecia que, prácticamente describe una línea recta, quitando un pico destacable en el día 12, pero la gráfica que describe se puede considerar estable.

Este hecho podría indicar que, a lo largo del periodo de obtención de tweets, los usuarios han cambiado poco de opinión sobre Ciudadanos, a pesar de la actividad política que se ha dado en el mismo.

6.3.3 Partido Popular

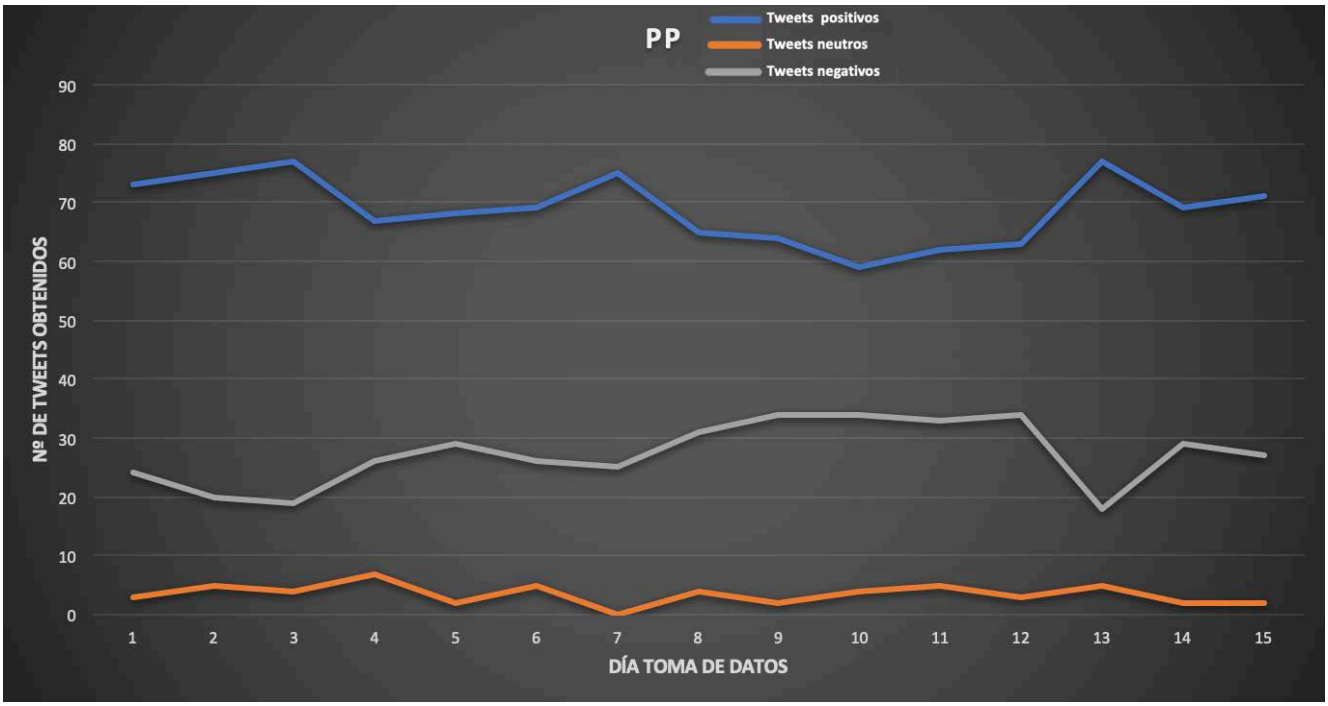


Figura 6.4 Evolución temporal de los tweets Partido Popular

Tweets totales	1496
Porcentaje de tweets positivos	1034 0,691176471
Porcentaje de tweets neutros	53 0,035427807
Porcentaje de tweets negativos	409 0,273395722

Tabla 6.5 Tabla resultados Partido Popular

Toma de datos	Día correspondiente
1	31-12-2018
2	01-01-2019
3	02-01-2019
4	03-01-2019
5	04-01-2019
6	05-01-2019
7	06-01-2019
8	07-01-2019
9	08-01-2019
10	27-01-2019
11	28-01-2019
12	29-01-2019
13	30-01-2019
14	31-01-2019
15	01-02-2019

Tabla 6.6 Correspondencia toma de datos Partido Popular

El Partido Popular tiene un total de 1496 tweets, los cuales, se han categorizado así:

- Se obtuvieron 1034 tweets como positivos, lo que se traduce a un 69,11% del total.
- Como tweets neutros se ha obtenido un 3.5% del total, 53 tweets de los 1496 totales.
- Como impresiones negativas se han categorizado 409 tweets, es decir, un 27,39%.

En la gráfica se observa un fenómeno semejante al que se daba en la gráfica de Ciudadanos, aunque esta, describe trayectorias menos estables, también presenta un rango de valores considerablemente continuo, con pocos picos, máximos o mínimos, en las trayectorias tanto positiva como negativa.

A pesar de lo descrito anteriormente, si se observa que las impresiones negativas tienen una progresión ascendente a lo largo del periodo de recolección de tweets, y al finalizar este, se presenta un pico que coincide con el valor menos negativo, el cual, coincide con el máximo más positivo del partido. Los valores positivos tienen una evolución inversa a los negativos, comenzando el periodo con los valores más altos y van disminuyendo a medida que transcurre el tiempo y los valores negativos aumentan.

6.3.4 Partido Socialista Obrero Español

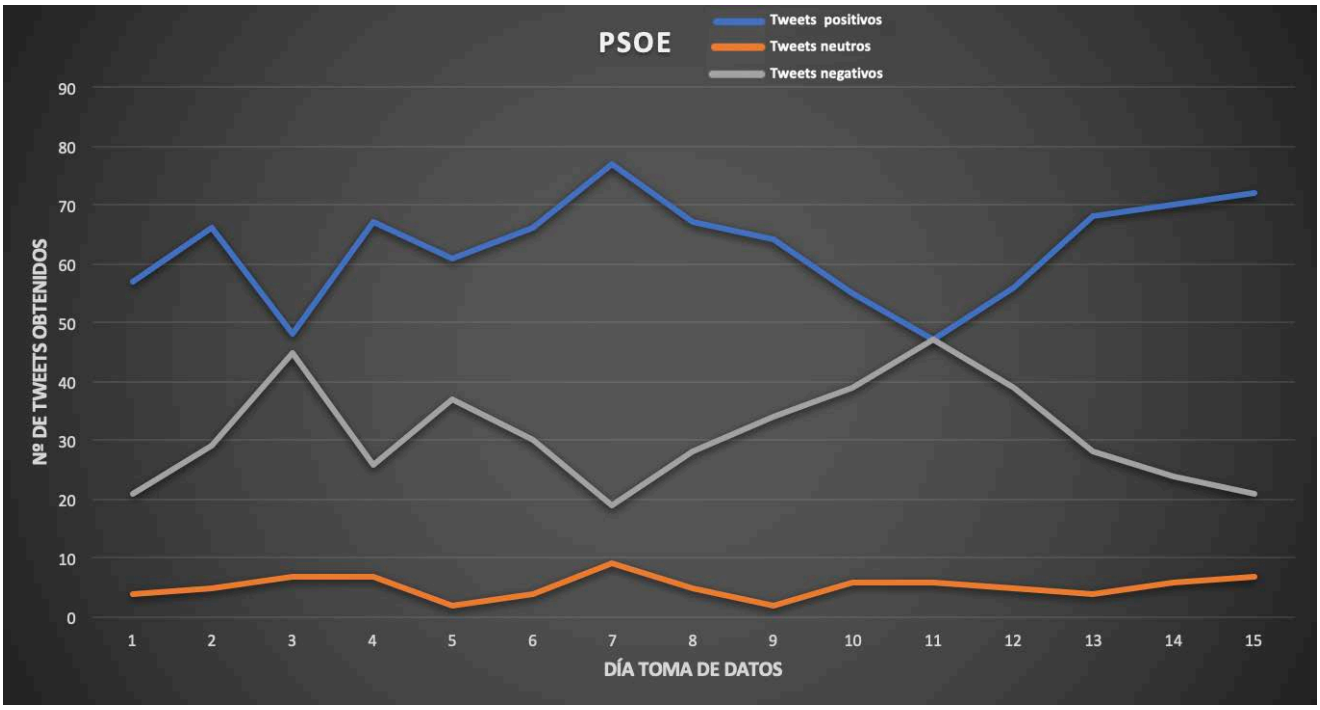


Figura 6.5 Evolución temporal de los tweets Partido Socialista

Tweets totales	1487
Porcentaje de tweets positivos	941 0,632817754
Porcentaje de tweets neutros	79 0,053127102
Porcentaje de tweets negativos	467 0,314055145

Tabla 6.7 Tabla resultados Partido Socialista

Toma de datos	Día correspondiente
1	01-01-2019
2	02-01-2019
3	03-01-2019
4	04-01-2019
5	05-01-2019
6	06-01-2019
7	07-01-2019
8	08-01-2019
9	09-01-2019
10	27-01-2019
11	28-01-2019
12	31-01-2019
13	01-02-2019
14	02-02-2019
15	03-02-2019

Tabla 6.8 Correspondencia toma de datos Partido Socialista

El Partido Socialista Obrero Español obtuvo 1437 tweets, de los cuales,

- 941 fueron considerados positivos por el sistema, un 63,28%,
- 79 neutros, 5,32% del total, y
- 467 negativos, que equivale a un 31,40%.

La gráfica de este partido es la más llamativa de todas, dado que, presenta muchos picos o cambios de valor a lo largo del periodo. Respecto a la forma, la gráfica es muy simétrica, teniendo en cuanta principalmente esos picos máximos y mínimos que se han comentado anteriormente.

El análisis de los datos obtenidos por el PSOE presenta varios fenómenos interesantes, como son, en la tercera medición, las curvas de las impresiones positivas y negativas llegan casi a coincidir, lo cual indica que, en ese día o periodo, las opiniones sobre el partido en la red estaban muy divididas.

Posteriormente aparecen picos significativos en los que las impresiones positivas aumentan, hasta tal punto que, llegan al máximo de la gráfica, pero seguidamente comienza un declive hasta llegar al mínimo de estas.

Para las impresiones negativas ocurre el mismo fenómeno, pero de forma inversa, lo cual es lógico, el número de impresiones negativas desciende hasta su mínimo, para después aumentar y ascender hasta el máximo.

Ambas impresiones, positivas y negativas, coinciden en el punto 11 del periodo de obtención, ambas toman el mismo valor, por lo que, el fenómeno anteriormente descrito se repite, las opiniones están muy divididas sobre el partido. Desde ese punto de encuentro entre ambas, las impresiones aumentan y las negativas por el contrario descienden de forma prolongada.

6.3.5 Partido Podemos

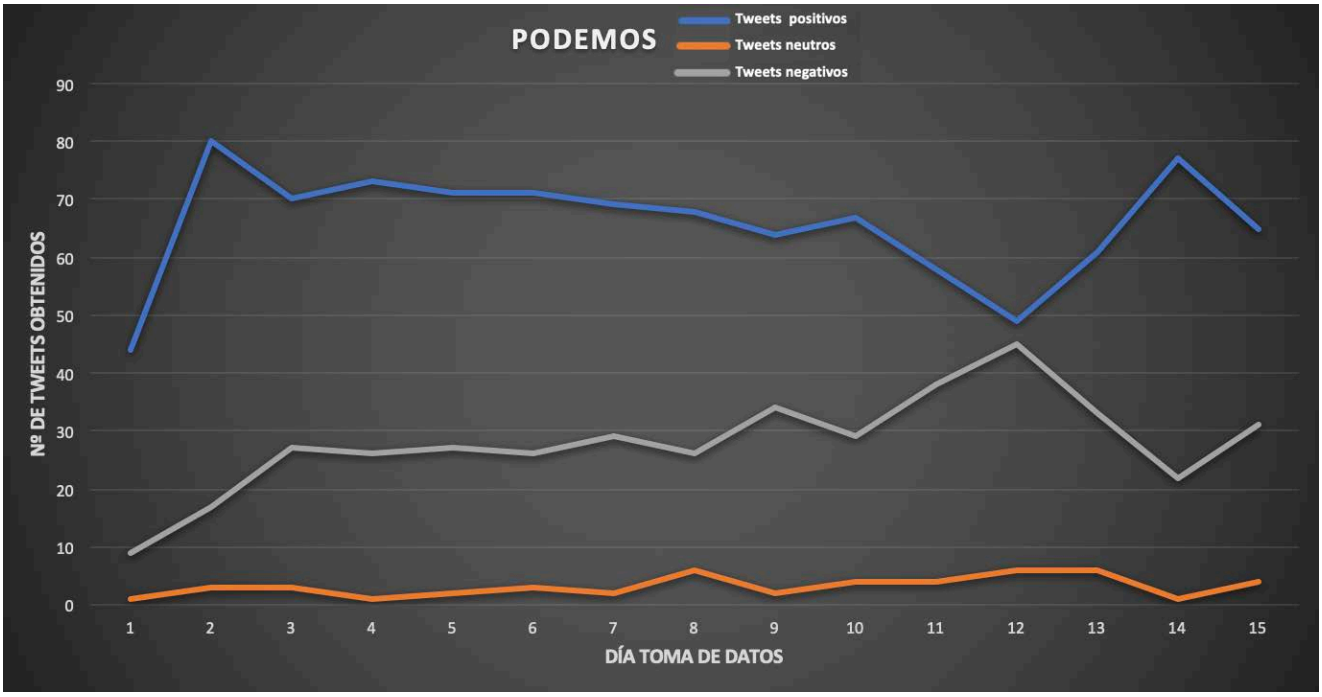


Figura 6.6 Evolución temporal de los tweets Podemos

Tweets totales	1454
Porcentaje de tweets positivos	987 0,678817056
Porcentaje de tweets neutros	48 0,03301238
Porcentaje de tweets negativos	419 0,288170564

Tabla 6.9 Tabla resultados Podemos

Toma de datos	Día correspondiente
1	01-01-2019
2	02-01-2019
3	05-01-2019
4	06-01-2019
5	07-01-2019
6	08-01-2019
7	09-01-2019
8	10-01-2019
9	11-01-2019
10	12-01-2019
11	18-01-2019
12	28-01-2019
13	31-01-2019
14	01-02-2019
15	02-02-2019

Tabla 6.10 Correspondencia toma de datos Podemos

Para el partido dirigido por Pablo Iglesias se analizaron un total de 1454 tweets, de los cuales:

- Se clasificaron 987 tweets como positivos, lo que se traduce a un 67,88%.
- De los tweets considerados como neutros son 48 tweets, lo que es un 3,32%.
- Los 419 restantes el sistema los ha considerado como negativos, lo que equivale a un 28,79%.

El principio de la gráfica presenta el valor más alto en cuanto a impresiones positivas se refiere. Posteriormente se describe un declive no muy prolongado, pero si continuo, a lo largo de toda la gráfica hasta llegar al punto 12, es decir, de 15 mediciones, en 11, los valores de las impresiones positivas descienden progresivamente. Después se incrementan hasta llegar a un valor semejante al máximo alcanzado al principio de la gráfica. Para finalizar con las impresiones positivas, pasado ese aumento, vuelven a descender. Ese descenso progresivo, se traduciría en un descontento general de los usuarios a lo largo de un periodo relativamente prolongado en el tiempo con el partido, lo cual tiene sentido, dado que los meses de diciembre, enero y febrero fueron negativos para la agrupación morada.

De forma inversa, las impresiones negativas comienzan en el punto más bajo de la gráfica y describe un aumento progresivo hasta llegar a la medición 12. A la vez que las impresiones positivas describen un aumento, las negativas sufren un pequeño declive, que queda relativamente lejos de su valor mínimo. Al finalizar aumentan de nuevo las impresiones negativas.

6.3.6 Comparación general de los resultados de los partidos políticos

A continuación, se muestran las gráficas generales con las impresiones de los partidos, tanto positivas como negativas. Además, se ha diseñado una tabla que recoge la información que se ha detallado anteriormente, los porcentajes de los tweets positivos y negativos.

Tanto las gráficas como la tabla se han diseñado con objeto de obtener una percepción global de los resultados obtenidos, siendo más sencillo establecer relaciones y comparaciones entre las impresiones de los usuarios.

En las gráficas cada partido está representado por el color con el que se le identifica.

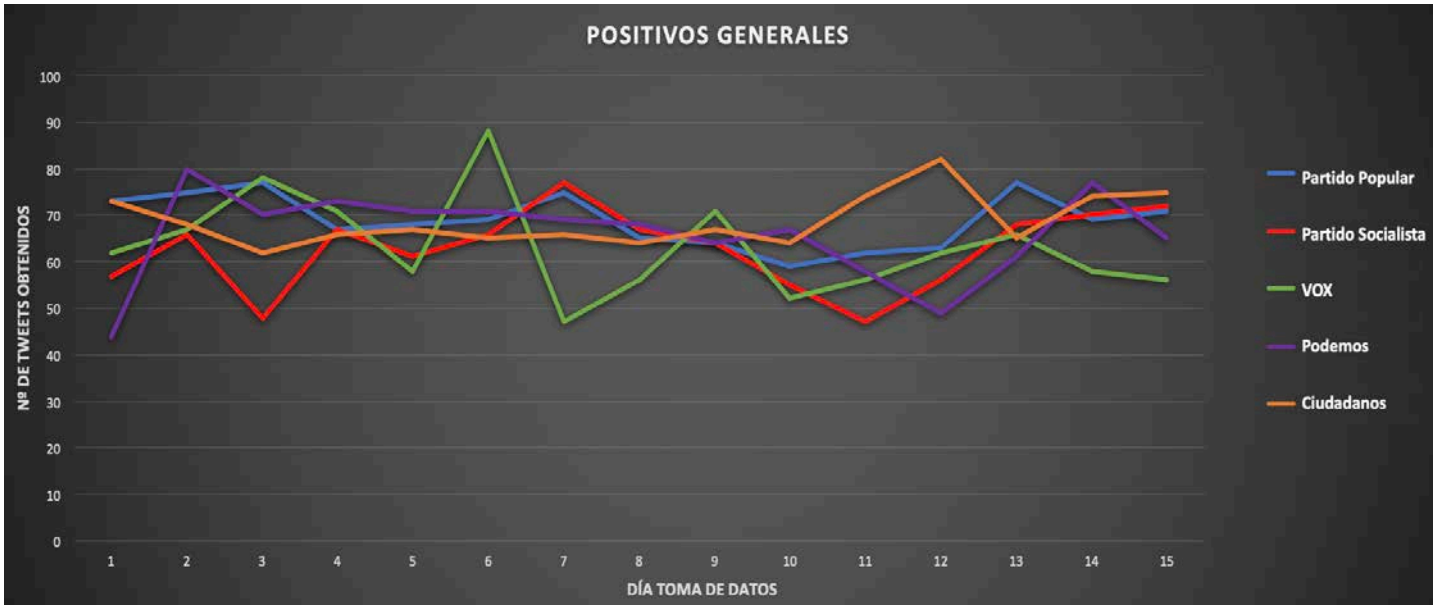


Figura 6.7 Evolución temporal de los tweets positivos generales

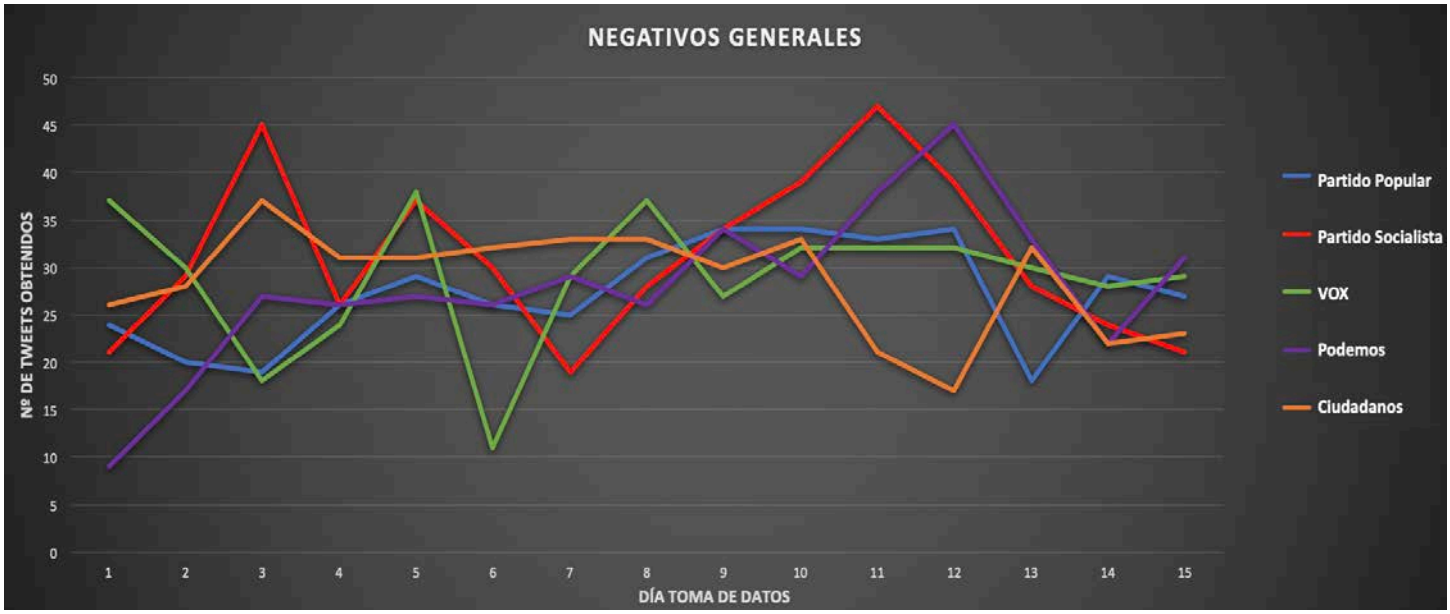


Figura 6.8 Evolución temporal de los tweets negativos generales

	Tweets positivos	Tweets negativos
VOX	948 0,661549197	434 0,30286113
CIUDADANOS	1032 0,688	429 0,286
PP	1034 0,691176471	409 0,273395722
PSOE	941 0,632817754	467 0,314055145
PODEMOS	987 0,678817056	419 0,288170564

Tabla 6.11 Tabla comparativa resultados generales

El partido que más impresiones positivas ha obtenido ha sido el Partido Popular, con un 69,11%, además de ser el que menos porcentaje de tweets negativos presenta, 27,39%. Esto arroja dos datos importantes, el primero que los resultados presentados por los datos son coherentes y el otro es la poca relevancia o importancia que tienen las opiniones consideradas como neutras, ya que, son muy pocas. De este dato o fenómeno se hablará de nuevo más adelante.

La línea que representa los tweets positivos del Partido Popular se mantiene en un rango de valores alto durante toda la gráfica, no de una forma destacable, pero si continua.

De forma opuesta, el partido con mayor porcentaje negativo es el Partido Socialista Obrero Español, 31,40%, el cual, describe una evolución opuesta a la del PP, ya que también es el menos positivo, con un 63,28%.

Observando las gráficas, en la primera, la positiva, se ve que describe una trayectoria semejante a la que describe Podemos, pero con más picos negativos. Si se observa la gráfica de negativos generales, la línea del PSOE destaca de forma considerable, dado que, describe varios picos, comentados anteriormente, que son los máximos de toda la gráfica, contando con todos los partidos.

Si en la gráfica de positivos generales, el PSOE iba parejo con Podemos, en esta gráfica se observa que en una mitad de la gráfica los sentimientos de los *tweets* obtenidos son semejantes pero la otra mitad no. Esto podría traducirse en que, la opinión sobre la formación morada tiene una progresión continua negativa, y las impresiones sobre los socialistas cambian de forma más brusca.

Además, el partido con mayor indecisión o, lo que es lo mismo, con mayor número de tweets clasificados como neutros también es el PSOE, con un 5,3%.

El hecho de que el partido más positivo sea el menos negativo, y, por el contrario, el menos positivo sea a la vez el más negativo, se debe a la poca relevancia que presentan las impresiones neutras, como ya se ha comentado anteriormente.

Se podía esperar que hubiese más dudas entre los usuarios, y eso, describiese situaciones distintas, como en la que VOX, dado que es un partido que ha resurgido en las elecciones autonómicas andaluzas y su posición entre los partidos es nueva, crease mayor incertidumbre o dudas, lo cual presentaría menos impresiones positivas y más neutras. De ser así, el PSOE seguiría siendo el partido más negativo, pero no el menos positivo.

Con esta situación hipotética se quiere mostrar que el hecho de ser el partido más positivo no tendría por qué traducirse de forma directa en ser el menos negativo o viceversa, sino que, hay más factores que influyen en estos resultados.

Este análisis deja como conclusión que, los usuarios de Twitter, mediante los tweets o textos que publican en sus cuentas dejan su opinión clara sobre cada partido por la poca relevancia que representan las impresiones neutras.

Una vez analizados los resultados numéricos de la experimentación, el foco del análisis pasa a las gráficas y las trayectorias que se describen.

En España, hasta la llegada de partidos como Ciudadanos y Podemos, gobernaba el bipartidismo: el Partido Popular y el Partido Socialista Obrero Español. Al llegar los nuevos partidos las opiniones de los usuarios se han dividido, y, siguen surgiendo más partidos, como VOX en Andalucía. Actualmente, según los expertos, los principales partidos se clasifican en la izquierda, PSOE y Podemos, y la derecha, PP, Ciudadanos y VOX.

Analizando la gráfica de generales positivos, los picos más importantes o destacables, los presentan partidos relacionados con la derecha política, VOX y Ciudadanos, el Partido Popular, por otro lado, es el partido más antiguo y más asentado en la política de estos tres, pero no aporta ningún pico importante entre sus impresiones, a pesar de este hecho, como ya se ha comentado mantiene sus valores positivos altos.

Por otra parte, se representa la izquierda política, PSOE y Podemos, los cuales describen los valores más bajos de esta gráfica, con dos picos destacables cada uno.

VOX, último partido en aparecer en el panorama político, por sus resultados en las últimas elecciones autonómicas andaluzas, describe una trayectoria llamativa. Como se ha mencionado anteriormente, presenta un pico con un valor muy positivo, pero también presenta un pico en valores inferiores, a la altura de los partidos de la izquierda.

Este fenómeno se podría deber a la gran controversia que ha generado entre la población española la batería de ideas de este partido. Dentro de los resultados obtenidos por los partidos que componen la derecha, VOX, es el menos positivo, que también se podría relacionar con lo anteriormente comentado.

Pasando a la gráfica de negativos generales, destacan a simple vista los picos descritos por Podemos y PSOE, que se relaciona directamente con los resultados obtenidos en las impresiones positivas. El PSOE presenta dos picos, los cuales son, máximos de la gráfica general. Podemos presenta el mínimo de la gráfica, lo que muestra su progresión negativa, dado que, después los valores aumentan.

La derecha política española muestra tres picos significativos, uno por cada uno de los partidos que lo componen, pero el más importante lo presenta VOX, directamente relacionado con el resultado tan positivo ya comentado en la anterior gráfica.

Como conclusión del análisis de las gráficas se podría obtener que, en el periodo de recolección, los partidos de la derecha han obtenido resultados más positivos que los partidos de la izquierda.

Este hecho se podría explicar por el cambio de gobierno en Andalucía, donde han pasado a gobernar el Partido Popular, Ciudadanos y VOX, lo cual es un hecho histórico, puesto que, durante los últimos 36 años había gobernado la izquierda, el PSOE, y con el mal momento general por el que atraviesan actualmente, o en los últimos meses, tanto Podemos como PSOE.

6.4 Relación de los resultados obtenidos con noticias políticas

A continuación, se relacionan algunos puntos importantes de las gráficas, principalmente los picos donde los valores han tomado cifras muy altas o muy bajas con noticias políticas que explicarían dichos valores.

El punto máximo que alcanza la gráfica de VOX, en su vertiente positiva, se recogió sobre la primera semana de enero, más concretamente el día 4, momento en el que el CIS preveía una subida importante en unas futuras elecciones generales, dado que, en dichas elecciones, pero del pasado 2016, VOX presentaba un 0,2% y, con esa subida pasaría a un 6,5%, lo que supondría una presencia destacada en el Congreso de los Diputados [15].

Además, a lo anterior se suma que Abascal y Casado comenzarían, por dichas fechas, acercar posturas para llegar acuerdos en Andalucía. El Partido Popular daría ayudas a hombres por violencia doméstica [16].

En Podemos, el aumento progresivo de forma negativa coge mayor velocidad en los puntos del eje horizontal, 11 y 12, llegando en este último a su valor máximo.

Estos puntos se refieren a los días que transcurren del 18 al 28 de enero, aproximadamente, más concretamente, del 18 y del 28 de enero. Sobre el día 18 se destaca el anuncio de la huida de Iñigo Errejón de Podemos, dejando así, su escaño como diputado, debido a las diferencias ideologías con Pablo Iglesias [17]. Poco más adelante, el día 28, se hace pública la pérdida de escaños por parte de Podemos, los cuales, pasan al poder del PSOE, debido a la huida de Errejón, lo que, además, desembocaría en un adelanto de la cumbre del partido.

En Ciudadanos, el ligero pico positivo que presenta es del día 28 de enero, momento en el que la formación morada dirigida por Pablo Iglesias queda por detrás del partido de Albert Rivera [18]. A pesar de ello, la gráfica seguidamente desciende, que coincidiría con la pequeña crisis del partido debido a la perdida de escaños que pasarían al Partido Popular.

En el Partido Socialista Obrero Español, se destacan dos picos negativos, que se dan en los puntos 3 y 11 del eje horizontal. El punto 3 consta del día 3 de enero, cuando la mitad de los votantes del partido rechazaban dialogar con Torra sobre el problema catalán y pedían a Pedro Sánchez que aplicase el articulo 155 [19]. El hecho de que tantos votantes del partido tuviesen dichas exigencias, explicaría que las impresiones positivas y negativas de los usuarios tuvieran en dicho punto valores muy parecidos.

Por otro lado, el punto 11, se corresponde con el día 28 de enero, cuando Sánchez no reconoció como legítimo el cambio de Guaidó por Maduro en la presidencia del gobierno venezolano. Dicha noticia tuvo mucha controversia dadas las circunstancias del país sudamericano y las reacciones instantáneas de otros países en reconocer ese cambio como legítimo [20]. La decisión del gobierno español no se comprendió, lo que explicaría las impresiones recogidas en la red social, donde el número de opiniones negativas y positivas son iguales.

El Partido Popular, dada la descripción de su gráfica y la búsqueda de acontecimientos en diversos medios, no se consideró ninguna relación relevante.

6.5 Análisis de resultados y conclusiones generales

Con lo expuesto anteriormente, se puede obtener la conclusión de que, acontecimientos de peso en la política española, se relacionan directamente con los resultados obtenidos mediante el análisis de la opinión o sentimiento de los tweets de los usuarios de dicha plataforma o red social. Es importante comentar que el ejemplo anterior muestra el análisis realizado durante un periodo concreto de tiempo; sin embargo, la finalidad principal de esta experimentación es mostrar cómo se obtiene el objetivo planteado en cualquier periodo de tiempo.

Además, mediante el análisis del sentimiento de los usuarios de una red social se podrían realizar otros estudios para entender el comportamiento de la sociedad. A parte del ámbito político se puede llevar a otros ámbitos, donde la opinión de las personas que forman la sociedad es clave, como, por ejemplo, detectar en qué gastar el dinero público, cuáles son las prioridades de los usuarios teniendo en cuenta de su edad y sexo o comprobar de qué manera afectan las acciones globales a la propia sociedad.

7 Gestión del proyecto

En el siguiente apartado se explicarán los puntos relacionados con la gestión del proyecto. Entre los puntos que se detallarán están la planificación del trabajo, el presupuesto estimado para llevar a cabo el proyecto y el impacto socioeconómico.

7.1 Planificación

El trabajo realizado se puede dividir en dos partes importantes, la recopilación y preparación de datos y la experimentación y estudio de los modelos obtenidos de la experimentación con el sistema.

Primeramente, se hizo un estudio y análisis sobre la actualidad de las herramientas que había disponibles para desarrollar las actividades planificadas, temas que tratar dentro del ámbito de la minería de texto y estudios relacionados con el presente trabajo, que sirviesen de guía para estructurar a este.

La tarea de recopilación comenzó a principios del mes de septiembre de 2018, más concretamente el día 14. En este punto las tareas que se realizaron básicamente fueron de recopilación de instancias de los distintos temas que se emplearían para los experimentos y establecer las relaciones y permisos entre las distintas partes del sistema. Además, una vez obtenido un número considerablemente grande de instancias se pasó a conformar los ficheros de distintas extensiones con los que se iba a experimentar. Esta tarea finalizó entre principios y mediados de febrero de 2019.

La segunda parte del trabajo se focaliza en la utilización de la herramienta para diseñar los distintos procesos que se querían ejecutar y, el estudio y análisis de los resultados alcanzados de los procesos realizados, a lo que se le suma la realización de la memoria del propio trabajo.

Además, durante ambas fases del proyecto se han realizado reuniones con el tutor del trabajo para realizar un seguimiento del estado de las tareas. Estas reuniones tenían como fin resolver los problemas encontrados de la manera más eficiente y rápida posible para que dichos problemas no penalizaran a la planificación del trabajo.

En la siguiente tabla se muestran las fechas de inicio y fin de las fases anteriormente descritas, con las horas aproximadas dedicadas a cada una.

	Fecha de inicio	Fecha de finalización	Horas dedicadas
Análisis y revisión del estado del arte (Fase 1)	7/09/2018	14/09/2018	12 horas
Obtención de datos (Fase 1)	14/09/2018	8/02/2019	22 horas
Configuración de herramientas externas (Fase 1)	9/02/2019	9/02/2019	2 horas
Preparación de ficheros (Fase 1)	10/02/2019	18/02/2019	17 horas
Experimentación y análisis de los procesos con el sistema (Fase 2)	19/02/2019	23/03/2019	41 horas
Memoria	25/03/2019	7/05/2019	120 horas

Tabla 7.1 Planificación fases principales del proyecto

A continuación, en la Tabla 7.2 se muestra la planificación de las tareas de las que está compuesta la memoria con las fechas de inicio y final de cada tarea, y las horas empleadas.

	Fecha de inicio	Fecha de finalización	Horas dedicadas
Introducción	25/03/2019	28/03/2019	13 horas
Estado del arte	29/03/2019	7/04/2019	27 horas
Análisis y diseño del sistema	8/04/2019	12/04/2019	12 horas
Resultados y evaluación del sistema	13/04/2019	30/04/2019	55 horas
Clasificación	13/04/2019	18/04/2019	23 horas
Agrupamiento	19/04/2019	22/04/2019	12 horas
Análisis del sentimiento	23/04/2019	30/04/2019	20 horas
Gestión del proyecto	1/05/2019	5/05/2019	10 horas
Conclusiones	6/05/2019	7/05/2019	3 horas

Tabla 7.2 Planificación tareas del proyecto

En la Figura 7.1 se presenta el cronograma que ha seguido el trabajo desde la recolección de los datos hasta la finalización de la memoria.

			Septiembre				Octubre				Noviembre				Diciembre				Enero				Febrero				Marzo				Abril				Mayo			
Nº Actividad	Inicio	Final	Semana 1	Semana 2	Semana 3	Semana 4	Semana 1	Semana 2	Semana3	Semana 4	Semana 1	Semana 2	Semana 3	Semana 4	Semana 1	Semana 2	Semana 3	Semana 4	Semana 1	Semana 2	Semana 3	Semana 4	Semana 1	Semana 2	Semana 3	Semana 4	Semana 1	Semana 2	Semana 3	Semana 4	Semana 1	Semana 2	Semana 3	Semana 4	Semana 1	Semana 2	Semana 3	Semana 4
Análisis y revision del estado del arte	7/9/18	14/9/18																																				
Obtención de datos	14/9/18	8/2/19																																				
Configuración con sistemas externos	9/2/19	9/2/19																																				
Preparación de ficheros	10/2/19	18/2/19																																				
Experimentación y análisis	19/2/19	23/3/19																																				
Introducción	25/3/19	28/3/19																																				
Estado del arte	29/3/19	7/4/19																																				
Análisis y diseño del sistema	8/4/19	12/4/19																																				
Clasificación	13/4/19	18/4/19																																				
Agrupamiento	19/4/19	22/4/19																																				
Análisis del sentimiento	23/4/19	30/4/19																																				
Gestión del proyecto	1/5/19	5/5/19																																				
Conclusiones	6/5/19	7/5/19																																				

Figura 7.1 Cronograma principal del proyecto

7.2 Presupuesto

En la presente sección se detallará el presupuesto necesario para realizar el proyecto que supondría este trabajo de fin de grado.

Los costes que se tendrán en cuenta serán: los costes de personal y de material.

7.2.1 Coste de personal y material

En el presente apartado se detallarán los costes correspondientes al personal y al material empleado para realizar el proyecto.

7.2.1.1 Coste de personal

Dado un proyecto, dentro del personal que lo realiza se distinguen varios roles.

- Jefe de proyecto: es el responsable de organizar, gestionar y dirigir el equipo de trabajo.
- Responsable configuración: es el encargado de definir, reconocer, proporcionar información y gestionar las modificaciones en la configuración del sistema, así como los cambios y versiones.
- Responsables análisis: es el encargado de realizar una descripción detallada del sistema, a través de un conjunto de requisitos y un grupo de modelos que satisfagan las necesidades de los usuarios para lo que se genera el sistema y que serán el origen para el proceso de diseño del sistema.
- Desarrollador: Persona encargada del desarrollo y comprobación el sistema.

En la Tabla 7.3, se realiza una estimación del coste/hora de cada miembro del equipo del proyecto, con el coste total de cada uno de ellos.

ROL	Coste por hora (€)	Horas totales de trabajo	Coste total (€)
Jefe de proyecto	29,74€	15 horas	446,10€
Responsable configuración	27,52€	26 horas	715,52€
Responsables análisis	27,52€	37 horas	1.018,24€
Desarrollador	20,87€	118 horas	2.462,66€
Totales	**---**	**196 horas**	**4.642,52€**

Tabla 7.3 Costes de personal del proyecto

7.2.1.2 Coste material

En esta sección, se va a realizar el desglose de los gastos destinados a la compra de los diferentes equipos electrónicos empleados para la correcta realización del proyecto. Además de los equipos que serían el hardware, también se añadirán los costes en software.

Para calcular los costes de los distintos elementos se tendrán en cuenta tres conceptos, el periodo de amortización, el coste del equipo o software y el tiempo de uso en el proyecto. El coste será el precio total del producto por el tiempo de uso entre el periodo de amortización.

A continuación, se detallarán los costes en la Tabla 7.4.

Producto	Precio total (€)	Unidades	Periodo de amortización	Tiempo de uso en proyecto	Coste para el proyecto
MacBook Pro 13.3”	1.300,00€	3	48 meses	4 meses	325,00€
Licencia Microsoft Office 2016	299,00€	1	48 meses	4 meses	24,92€
Licencia *RapidMiner* Studio	5.000,00€	1	12 meses	4 meses	1666,67€
Total	**---**	**---**	**---**	**---**	**2016,59€**

Tabla 7.4 Costes materiales del proyecto

7.2.2 Coste total

A los costes materiales y costes de personal calculados anteriormente, se le tienen que sumar otra serie de costes que vienen dados por: el beneficio de realizar el proyecto, los impuestos y un margen para imprevistos o de riesgo.

- Margen para imprevistos o de riesgo: se considera un 10 %, para posibles contratiempos materiales.
- Beneficio del proyecto: se considera un 30 %, lo cual sería compensación para la empresa por realizar el proyecto.
- Impuestos: se añadirá un 21% del coste generado hasta el momento, que representará el IVA.

En la Tabla 7.5 se muestra el cálculo de los costes descritos.

Concepto	Cantidad (€)
Costes de personal	4.642,52€
Costes de material	2.016,59€
Total	6.659,11€
Margen de riesgo	665,92€
Beneficio del proyecto	1997,74€
Base imponible	9.322,77€
Impuestos (IVA)	1.957,79€
Total del proyecto	11.280,56€

Tabla 7.5 Costes totales del proyecto

7.3 Impacto socioeconómico

Los objetivos principales del proyecto que se desarrollan en este trabajo son la presentación de un sistema de clasificación fiable y el empleo del análisis de sentimiento como herramienta válida para extraer información de la sociedad en la que vivimos actualmente.

El impacto socioeconómico esperado al aplicar ambos puntos se espera que cubra diferentes ámbitos al ser dos objetivos relacionados, pero claramente diferentes.

Por parte del sistema de clasificación, que se pueda emplear el mismo para la clasificación de otro tipo de textos o datos con el fin de obtener conclusiones y ver cómo se relacionan los datos entre sí. Aplicado a una empresa, podría solucionar problemas para estructurar las tareas en función de los ámbitos que cubre cada departamento basándose en la temática de la propia tarea, lo que sería un distribuidor de tareas. Además de poder aplicarse al entorno funcional de la empresa, también podría emplearse para analizar los propios datos generados de la empresa y ver, en caso de que estos datos presenten relaciones, cómo y por qué se relacionan.

Ambas aplicaciones darían ventajas competitivas respecto a sus competidoras, puesto que, se mejoraría el funcionamiento interno de la misma y analizar los datos generados, retroalimentarían a la propia empresa y ayudaría en la hora de tomar de decisiones.

Respecto el análisis de sentimiento, se considera que podría ser una herramienta potencialmente interesante como sustituto o complemento a las encuestas y sondeos que

ya se conocen. Una herramienta así, permitiría analizar una cantidad de opiniones muy grande con muy poco coste, además de que están extraídas directamente de la sociedad, mediante las tecnologías que más se utilizan a día de hoy como son las redes sociales. Este análisis de sentimiento presenta ventajas importantes respecto a los sondeos y encuestas, dado que, permite obtener la opinión de los usuarios, de la sociedad de forma instantánea, y, además, se puede acotar la propia búsqueda de opinión. Esto último es importante porque permite realizar estudios de las opiniones dado un evento concreto en ese mismo instante de forma inmediata o recopilar información a lo largo del tiempo y realizar un estudio más general. Por la forma de recopilar los datos, el sistema permite mantener en el anonimato las opiniones o información que se analiza igual que las otras herramientas.

Por otro lado, este sistema aplicado al ámbito empresarial y de negocios permite influir en la sociedad para que la empresa obtenga ventajas competitivas frente a sus competidores del sector, consiguiendo una imagen de marca por encima de los mismos, además de otras ventajas.

8 Conclusiones

En esta sección se presentarán las diferentes conclusiones que se han extraído al realizar este trabajo, tanto técnicas como personales. Además, se expondrán los trabajos futuros que se podrían realizar para como continuación del presente documento.

8.1 Conclusiones

En este trabajo se han realizado distintos estudios mediante el análisis de los datos obtenidos desde Twitter.

El primero de los estudios ha consistido en analizar un conjunto de clasificadores para ver la evolución de cada uno de ellos con distintos conjuntos de datos conformados con textos de diferentes temas. El principal objetivo de este estudio ha sido obtener un clasificador de confianza.

Se analizaron cuatro clasificadores diferentes, *Naive Bayes*, *Deep Learning*, *KNN* y *Decision Tree*, mediante validación cruzada, donde se obtuvieron resultados significativos. *KNN* y *Decision Tree*, quedaron atrás, dado que los resultados que arrojaron, no eran lo suficientemente positivos para pasar a la siguiente evaluación. *Naive Bayes y Deep Learning*, por otra parte, pasaron a un segundo estudio donde se aplicaron los modelos para ver su evolución en una clasificación directa. Ambos obtuvieron resultados muy consistentes y positivos, pero el mejor clasificador fue *Naive Bayes*.

Además, tanto en la experimentación de clasificación con los mejores modelos como en la validación cruzada se observaron varios resultados importantes. Se esperaba que el hecho de incrementar el número de instancias de un conjunto de datos a otro, tuviese un efecto positivo en el proceso de aprendizaje de los clasificadores, pero se obtuvo el efecto contrario en muchos de los casos. Se obtuvo una relación entre los resultados de la clasificación de dos clases en prácticamente todos los experimentos. Esta relación cobra sentido ya que, los clasificadores confundían textos económicos como si fuesen textos políticos, y ambos temas están estrechamente relacionados en la sociedad en la que vivimos.

El segundo de los estudios ha consistido en realizar un agrupamiento mediante aprendizaje no supervisado con el algoritmo K-medias con los mismos conjuntos de datos con los que se realizó la clasificación. Los resultados en este caso, no fueron tan buenos como se esperaban, ya que, se esperaba un agrupamiento con cuatro *clusters* relativamente homogéneos en cuanto a cantidad de instancias se trata. En cambio, se obtuvo un *cluster* que abarcaba la gran mayoría de tweets de los conjuntos y tres agrupaciones casi vacías, sin apenas instancias.

El tercer y último estudio se relaciona con el análisis de sentimiento de los tweets de los usuarios respecto a la situación política del país. A lo largo de más de un mes se

descargaron textos vinculados a los principales partidos políticos de nuestro país como son: Partido Popular, Partido Socialista Obrero Español, Podemos, Ciudadanos y VOX.

El principal objetivo de este estudio era mostrar como una herramienta más este tipo de análisis de cara a sustituir o complementar la información de los sondeos y encuestas. Además, se estudió la polaridad de los textos, clasificándolos en positivos, negativos y neutros.

Los resultados obtenidos se estudiaron de forma global viendo la evolución de las opiniones de los usuarios a lo largo del tiempo recogiendo cambios de comportamiento o datos llamativos, así con cada partido.

Para concluir si el análisis planteado podía aportar la información planteada, esos datos destacados y cambios de opinión entre los usuarios se intentaron vincular a eventos políticos ocurridos en el mismo día de la obtención de los datos o días seguidamente posteriores. Al realizar esta tarea de vinculación se puede concluir que, en muchos de los casos, esos cambios de opinión entre los usuarios o datos destacados tenían el respaldo de un evento o noticia política importante.

Para finalizar con la conclusión y de forma más general, se considera que se han aplicado distintos conocimientos y herramientas relacionadas con la mención y los conocimientos obtenidos en el grado. El tema principal del trabajo y los objetivos planteados podrían tener una aplicación real, como las expuestas en la sección 7.3.

Además, se considera que, por lo concluido en esta sección, los objetivos principales planteados para este trabajo se podrían considerar como cumplidos.

8.2 Trabajos Futuros

A continuación, se muestran posibles proyectos relacionados con el desarrollado en este trabajo.

Respecto a la parte de las tareas de clasificación y agrupamiento:

- Realizar la validación cruzada con los mismos clasificadores que se han empleado en el presente trabajo con otro equipo y otros conjuntos de instancias. Al emplear otro equipo con mayor potencia, y ficheros con más instancias, se pretendería obtener mejores resultados con los clasificadores.

- Realizar la validación cruzada con otros clasificadores, con el fin de ver si sería posible obtener un clasificador con mejores resultados que el obtenido en este trabajo.

- Realizar un estudio semejante a este con temas relacionados entre sí, para analizar con mayor detalle el aprendizaje y evolución de los clasificadores al clasificar

textos con temas conectados entre sí. Se buscarían temáticas que formasen parte de uno más general.

- Repetir el experimento de agrupamiento con otros conjuntos de datos y otro equipo con el fin de conseguir resultados más concluyentes que los obtenidos en el presente trabajo.

Respecto a la parte de análisis de sentimiento:

- Realizar un estudio semejante al realizado en un periodo de pre elecciones generales, con el fin de analizar la opinión de la sociedad en un evento político de mucha relevancia.

- Realizar un estudio semejante al realizado en un momento en el que la sociedad viva cualquier situación con un impacto importante sobre la misma.

- Realizar un estudio semejante durante un periodo de tiempo prolongado, una legislatura, con el fin de ver la evolución de las opiniones de la sociedad según actúen los diferentes partidos políticos.

- Estudiar los rangos de edad y situación geográfica de los usuarios en función de sus opiniones sobre un partido o personaje político concreto.

9 Bibliografía

[1] K. Smith, «Brandwatch,» 2016. [En línea]. Available: https://www.brandwatch.com/es/2016/06/44-estadisticas-twitter-2016/.

[2] I. RapidMiner, «Rapid Miner Studio,» 11 03 2019. [En línea]. Available: https://rapidminer.com/get-started/.

[3] BSA, «data.bsa.org,» [En línea]. Available: https://data.bsa.org/wp-content/uploads/2015/10/BSADataStudy_es.pdf.

[4] J. Mejia, «www.juancmejia.com,» [En línea]. Available: https://www.juancmejia.com/marketing-digital/estadisticas-de-redes-sociales-usuarios-de-facebook-instagram-linkedin-twitter-whatsapp-y-otros-infografia/.

[5] ITelligent, «https://itelligent.es/es,» [En línea]. Available: https://itelligent.es/es/10-ventajas-la-mineria-web/.

[6] A. d. l. Peña, «resencia en Twitter de los candidatos a las elecciones madrileñas de 2015».

[7] G. Martínez, «Minería de datos. Cómo hallar una aguja en un pajar,» 2001.

[8] C. s. p. c. t. e. Twitter, «eprints.ucm,» [En línea]. Available: https://eprints.ucm.es/44660/1/Memoria%20TFG.pdf.

[9] C. a. d. t. p. e. s. d. c. e. e. r. sociales, «e-archivo.uc3m.es,» [En línea]. Available: https://e-archivo.uc3m.es/handle/10016/23769#preview).

[10] C. A. D. T. S. E. T. P. E. A. A. D. C. SUPERVISADA, «addi.ehu.es,» [En línea]. Available: https://addi.ehu.es/bitstream/handle/10810/22632/Tesis%20Final%20Oscar.pdf?sequence=1&isAllowed=y).

[11] P. d. t. p. p. T. E. A. 2012, «idus.us.es,» [En línea]. Available: https://idus.us.es/xmlui/handle/11441/66738.

[12] S. a. p. l. c. d. l. o. p. g. e. Twitter, «http://www.rcs.cic.ipn.mx,» [En línea]. Available: http://www.rcs.cic.ipn.mx/rcs/2015_95/Sistema%20automatico%20para%20la%20clasificacion%20de%20la%20opinion%20publica%20generada%20en%20Twitter.pdf.

[13] A. Ltd., «AYLIEN,» 11 03 2019. [En línea]. Available: https://aylien.com/.

[14] E. E. P. S.L., «EL PAIS,» 2 Diciembre 2018. [En línea]. Available: https://resultados.elpais.com/elecciones/2018/autonomicas/01/index.html.

[15] ©. 2. E. Press, «Europa Press,» 4 Enero 2019. [En línea]. Available: https://www.europapress.es/nacional/noticia-portadas-periodicos-viernes-enero-2019-20190104001619.html.

[16] C. d. R. y. T. Española, «RTVE,» 4 Enero 2019. [En línea]. Available: http://www.rtve.es/noticias/20190104/pp-ofrece-vox-ayudas-hombres-victimas-violencia-domestica/1863521.shtml.

[17] ©. 2. E. Press, «Europa Press,» 18 Enero 2019. [En línea]. Available: https://www.europapress.es/nacional/noticia-portadas-periodicos-viernes-18-enero-2019-20190118001101.html.

[18] ©. 2. E. Press, «Europa Press,» 28 Enero 2019. [En línea]. Available: https://www.europapress.es/nacional/noticia-portadas-periodicos-lunes-28-enero-2019-20190127235222.html.

[19] ©. 2. E. Press, «Europa Press,» [En línea]. Available: https://www.europapress.es/nacional/noticia-portadas-periodicos-jueves-enero-2019-20190103001522.html.

[20] ©. 2. E. Press, «Europa Press,» 28 Enero 2019. [En línea]. Available: https://www.europapress.es/nacional/noticia-portadas-periodicos-lunes-28-enero-2019-20190127235222.html.

[21] B.-E. Erlandsson, A. Dragomir y A. Akay, «Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care,» 2014.

[22] A. R. S. R. F. S. V. S. Preslav Nakov, SemEval-2016 Task 4: Sentiment Analysis in Twitter, Qatar, USA, 2016.

[23] P. N. S. K. V. S. A. R. S. M. M. Sara Rosenthal, SemEval-2015 Task 10: Sentiment Analysis in Twitter, 2015.

[24] BSADataStudy_es.pdf, «data.bsa.org,» [En línea]. Available: https://data.bsa.org/wp-content/uploads/2015/10/BSADataStudy_es.pdf.